Implementing assessment innovations in higher education

Boevé, Anna Jannetje

Implementing Assessment Innovations in Higher Education

Anna Jannetje Boevé

ico

Interuniversity Center for Educational Research

Title: Implementing Assessment Innovations in Higher Education

Printed: Ipskamp printing

Design: Wolf Art Studio, Jeroen van Leusden

Cover design: Wolf Art Studio, Jeroen van Leusden

ISBN: 978-94-034-0642-8

The research presented in this thesis was funded by the innovation budget of the Faculty of Behavioural and Social Sciences of the University of Groningen.

rijksuniversiteit
groningen

Implementing Assessment Innovations in Higher Education

Proefschrift

ter verkrijging van de graad van doctor aan de
Rijksuniversiteit Groningen
op gezag van de
rector magnificus prof. dr. E. Sterken
en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op

maandag 14 mei 2018 om 12.45 uur

door

Anna Jannetje Boevé

geboren op 9 oktober 1988
te Houten

Promotores
Prof. dr. R.R. Meijer
Prof. dr. R.J. Bosker

Copromotor
Dr. C.J. Albers

Beoordelingscommissie
Prof. dr. J. Cohen-Schotanus
Prof. dr. W.H.A. Hofman
Prof. dr. B.P. Veldkamp

Dankwoord

Bedankt voor jullie geduld, feedback en resultaatgerichtheid:

Rob, Roel en Casper. Met jullie hulp is dit proefschrift er gekomen zoals het er nu is. Ik heb veel van jullie geleerd dat ik meeneem in mijn verdere weg in de wetenschap.

Bedankt voor het mede mogelijk maken van de onderzoeken op dit proefschrift:

Rink Hoekstra, Jorien Vugteveen, Edith van Krimpen-Stoop, Mark Nieuwenstein, Berry Wijers, Yta Beetsma, Carlien Vermue, Hans Beldhuis en Jorge Tendeiro. Ik vond het een eer om bij jullie cursussen betrokken te zijn, data te mogen verzamelen en samen te werken aan verschillende onderzoeken in dit proefschrift.

Bedankt voor de gezelligheid en de koffie-/rookpauzes:

Susan, Edith, Karin, Jasperina en alle andere PhD studenten en collega's van Psychometrie en Statistiek bij de RUG. Door jullie ging ik met plezier naar mijn werk en ik vond het erg leuk om jullie te leren kennen en een paar jaar samen op te trekken.

Bedankt voor inspiratie en het delen van passie voor de wetenschap en onderwijs, voor steun en vriendschap:

Edith de Leeuw, Joop Hox, Peter Lugtig, Larike Bronkhorst, Wout Koelewijn, Max Aangenendt, Catherine Evers, Alexandra Dingemans, Astrid Junghans, Jan Durk Tuinier, Edsci-groep: Nienke, Lindy, Jolien, Bram en Joost, Edsci-nerdiesgroep: Raisa, Heleen, Esther, Sietske, Marijn, Suzanne en Bobby. Door jullie kon ik mijn passie voor onderwijs en wetenschap ontwikkelen, sparren over PhD perikelen en ondanks alles, ontdekken dat ik die passie voor onderwijs en wetenschap nog altijd met mij meedraag.

Bedankt voor het thuis kunnen zijn in Groningen toen ik daar woonde:

Annelies K, Loes, Mariska, Jeanette, Petrie, alle cantorij leden van de Nieuwe Kerk en dames van de Nieuwe Kerk. Door jullie werd mijn tijd in Groningen meer dan een PhD, het werd ook een dierbare periode in mijn leven.

Bedankt voor ontspanning, kletsen, gezelligheid, ontlading, afleiding, uithuilen en knuffels:

lieve paranimfen Floor en Hanneke, Hannah G, Emmaly en Hans, Frances, Annelies O, Annelies K, Laurien en Wilmar, Sifra en Thijs Willem, Jonne, Daphne en Jeroen, Marloes en Giel, vriendengroep uit Lunteren, kring van de Maranathakerk en de Libanon-groep. Door jullie kon ik blijven lachen, voelde ik mij niet alleen en werd ik er telkens aan herinnerd dat het leven meer omvat dan een PhD.

Bedankt voor jullie steun en eindeloze liefde:

Pa, Ma, Tim, Petra en Kees en schoonfamilie Peter, Ria en José. Door jullie kon ik altijd weer even aarden en dat gaf me kracht om door te zetten als ik er doorheen zat.

Bedankt voor je zorgzaamheid, vertrouwen en liefde Alex. Evenals voor je geduld met mijn verstrooidheid, tranen en spontaniteit. Trouwen met jou tijdens mijn PhD is de beste beslissing ooit geweest. Samen met jou kon ik het aan en kunnen we vooruitkijken naar nieuwe avonturen.

Contents

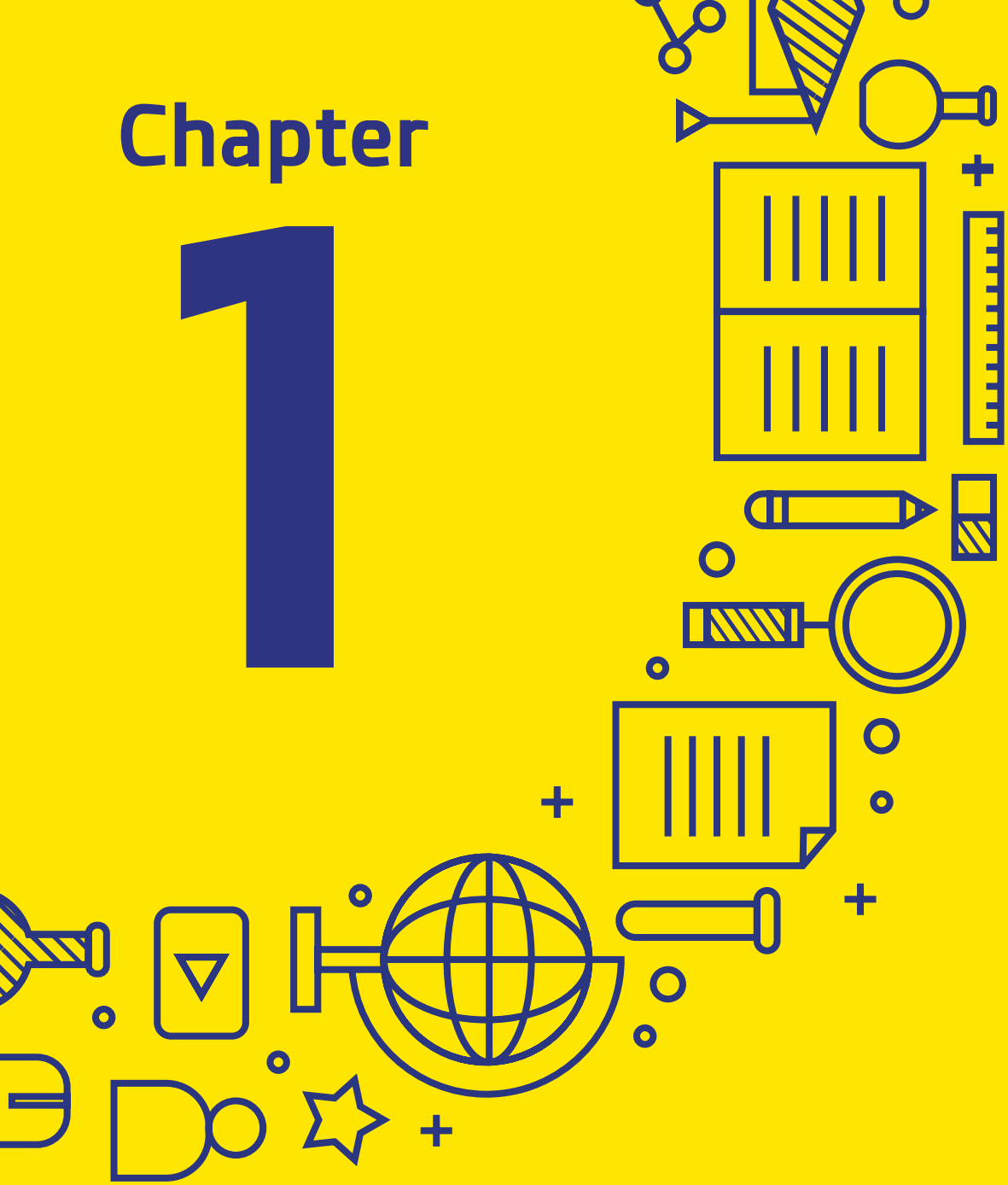
Chapter 5 Implementing the flipped classroom

Chapter 6 Natural variation in grades in higher education

Chapter 7 Discussion

Chapter

1



Introduction

1.1 Introduction

Student assessment plays an important role in higher education. Two recent developments in educational assessment were important for the research conducted in this thesis: (1) the massification of higher education and (2) the digital developments in higher education. Due to the massification of higher education, teachers are faced with large classes of, sometimes, hundreds of students (Hornsby & Osman, 2014). As a result the teacher-student ratio becomes very small, leaving teachers with little time and resources to monitor their students' learning process. As a result of the digital developments, it is not possible for students in the Netherlands to access university courses without having access to the Internet and computer or smart devices: from enrolment to course participation to the access of final grades, virtually all information is accessible and stored online. Although there are important new technological developments, students are often still required to tick the appropriate box on a paper-and-pencil answer sheet when completing exams. Thus the challenge facing universities is how to integrate digital technologies in a way that contributes to improving learning and assessment.

Another important aspect that inspired the research conducted in this thesis was the new system of performance-based funding that was recently introduced in the Netherlands in the form of an agreement between each university and the government (de Boer et al., 2015). Performance based funding of universities is a policy change that has occurred in the past few decades in countries across the world in different ways (De Boer et al., 2015). In the Netherlands, the agreement is that if after a certain period, the objectives stipulated in the agreement between a university and government have not been met that university will receive less funding. Important performance indicators agreed upon were the student dropout rates in the first year and the 4-year bachelor graduation rate. Although recognizing that maintaining quality in higher education should not be limited to these two indicators (Bussemaker, 2014), other indicators in the agreement remained vague. In line with the digital revolution, the minister of education also noted that: "New developments like open online education provide the opportunity to further improve the quality of higher education. Higher education should give the right attention to all these (existing and new) challenges" (Bussemaker, 2014).

In addition to the introduction of performance-based funding, the Dutch Inspectorate of Education recently evaluated the quality of assessment in higher education (Inspectie van het Onderwijs, 2016). One of the recommendations was to increase research and evaluation concerning assessment quality in higher education. This thesis is a collection of studies investigating the implementation of innovations in assessment in higher education.

1.1.1 Assessment

Assessment in education refers to the entire process of collecting information concerning students' knowledge, skills, and/or abilities (Cizek, 2009). A test or exam is a "systematic sample of a person's knowledge, skill or ability" (Cizek, 2009, pp.64). Test results can be used for different purposes, such as determining the strengths and weaknesses of students, guiding instruction, and making decisions about students (Cizek, 2009). A test is therefore a potential part of assessment, but assessment does not necessarily include tests. In this

thesis, some chapters are more focused on exams, and exam results (chapters 2, 3, and 6), while other chapters are more broadly related to assessment throughout a course (chapters 4 and 5).

An important distinction can be made between summative and formative functions of tests (Black & Wiliam, 2003; Wiliam & Black, 1996). The aim of a summative assessment is to evaluate student learning after instruction by comparing it to some standard, whereas the aim of formative assessment is to modify teaching and learning activities to improve the learning process. Examples of summative tests are final exams used to decide whether a student has sufficiently achieved the learning goals, or admission tests used to determine whether a candidate is sufficiently skilled to enter a particular education program. Examples of formative tests are diagnostic tests that aim to map parts of the literature that a student does not master yet. Thus, formative tests are assessments *for* learning rather than assessment *of* learning (Schuwirth & Van Der Vleuten, 2011). In this thesis, the focus in chapters 2 and 6 is on summative assessment. Often in educational practice, the distinction between formative and summative assessment is not always clear. This is a challenge for teachers in the implementation of assessment innovations as will be illustrated in chapters 3, 4, and 5.

Another, though subtler, distinction in the literature is between large-scale testing and classroom testing. In the Netherlands, examples of large scale testing are the national end-of-primary school tests (delivered by various organizations, e.g., CITO), and the national exams at the end of secondary education. In the US, a well-known example of large scale high-stakes testing is the SAT that is administered at the end of high school and used by colleges to select students. Other large-scale tests are admission tests such as the Law School Admissions Test (LSAT), and the Graduate Record Examinations (GRE). Large-scale tests are subject to very stringent quality criteria, requiring large amounts of resources. Furthermore, given the scale of testing, advanced statistical methods have been developed to evaluate the characteristics and quality of these tests and test items. Classroom testing, on the other hand, occurs in the classroom in the context of a learning process. The aim of classroom testing is to facilitate and evaluate the learning process of students. Tests used in the classroom are selected or developed by teachers, on a much smaller scale, and teachers generally do not have the resources, time, or amount of students to extensively evaluate the quality of tests. There is a growing interest in combining the strengths of both large-scale - and classroom testing, as a first convention on classroom testing was organized by the National Council on Measurement in Education in the US in 2017. In this thesis, the focus is on classroom testing, but statistical techniques developed in the context of large-scale testing are used in chapters 3 and 4.

1.1.2 The context of the research conducted in this thesis

All the studies in this dissertation were conducted at the University of Groningen, the Netherlands, and most chapters considered courses in the first year of the bachelor program. Study results in the first year are important because they have strong predictive validity for later study success (e.g., Niessen, Meijer, & Tendeiro, 2016) and because there is a binding study advice (BSA, in Dutch: bindend studie advies), that is, Dutch law requires

bachelor degree programs to inform students at the end of the first year whether they are allowed to continue their study. In practice, this advice is translated into rules about the number of credits students must minimally obtain; for example, some universities require that students have to obtain 45 out of the 60 European Credit Transfer and Accumulation System (ECTS) points at the end of their first year. If students do not attain sufficient ECTS points, they are not allowed to continue their study. In the Netherlands there is also a financial incentive for students to perform well: If they decide to quit after January tuition fees are not reimbursed. As a result of the binding study advice, in the first year of university education the stakes are high for students. Most of the first year courses are assessed by means of a final exam, largely consisting of multiple choice questions. The studies in this thesis were conducted in collaboration with teachers seeking to improve or innovate their assessment by implementing changes, often by means of technology.

1.2 Short introduction to each chapter of the thesis

This dissertation is organized as follows: In chapter 2 the implementation of computer-based exams is studied. Digital testing such as computer-based or web-based testing has the potential to ease and improve assessment in higher education. Nevertheless, there have been concerns about equality of test-modes, the fairness, and the stress students might experience (e.g., Whitelock, 2009). In order to ensure a smooth transition from traditional paper-based exams to computer-based exams in higher education, it is important that students perform equally well on computer-based and paper-based administered exams. If, for example, computer-based administration would result in consistently lower scores than paper-based administration, due to unfamiliarity with the test mode or due to technical problems this would result in biased measurement. Thus, it is important that sources of error, or construct irrelevant variance (Huff & Sireci, 2001), which may occur as a result of administration mode, are prevented or minimized as much as possible in high-stakes exams. The main research questions guiding this chapter were: How do students experience computer-based exams? And, is there a difference in student performance depending on mode of examination or preference for mode of examination?

In Chapter 3 the question was addressed whether and when reporting sub-test scores (or subscores) of assessment in higher education is useful. Given the limited time and resources teachers in higher education have given the large classes, teachers and management are often interested in efficient ways of giving students diagnostic feedback. Providing information on the basis of subscores is one method that is often used in large-scale standardized testing, and is more and more often under consideration in classroom testing in higher education as well. After a discussion of recent psychometric literature that warns against the use of subscores in addition to the use of total scores, I illustrate how the added value of subscores can be evaluated using two college exams: A multiple choice exam and a combined open-ended question and multiple choice exam. These formats are often used in higher education and represent cases in which using subscores may be informative.

In chapter 4 the focus is on implementing practice tests. There is a wealth of research on the positive effect of assessment on learning (e.g., Roediger & Karpicke, 2006), and the use of formative assessment to improve student learning is generally recommended. Nevertheless, it is unclear how to most effectively implement practice tests in college psychology education. In chapter 4 students' use of practice test resources was explored and evaluated in light of student performance. First, the relationship between student use of practice test resources and exam results was investigated in three courses with different types of implementations of practice tests. In a follow-up study for one of the courses, the performance of cohorts with practice test resources was compared to a cohort without practice test resources by means of test equating, a technique developed in the context of large-scale testing.

In chapter 5 the focus is on implementing the flipped classroom (e.g., Abeysekera & Dawson, 2015; Street, Gilliland, McNeil, & Royal, 2015). The flipped classroom is becoming more popular as a means to support student learning in higher education by requiring students to prepare before lectures and actively engage in the lectures. While some research has been conducted with respect to student performance in the flipped classroom, not much is known about students' study behaviour throughout a flipped course. Students' study behaviour throughout a flipped and regular course was explored in chapter 5 by means of bi-weekly diaries. Furthermore, student references to their learning regulation and study behaviour were explored in the course evaluations.

Chapter 6 was inspired by the research conducted in chapters 2 through 5 and is more methodologically oriented. To investigate the effect of innovations in the teaching-learning environment, researchers often compare student performance from different cohorts over time, or from similarly designed courses in the same academic year. However, it is important to recognize that variance in student performance can be attributed to both random fluctuation and to various innovations in higher education. Therefore, it is important to take the natural variation in student performance into account. The main question addressed in chapter 6 was: to what extent does student performance in first year courses vary within time, over time, and between courses and how can this information be used to evaluate educational innovations?

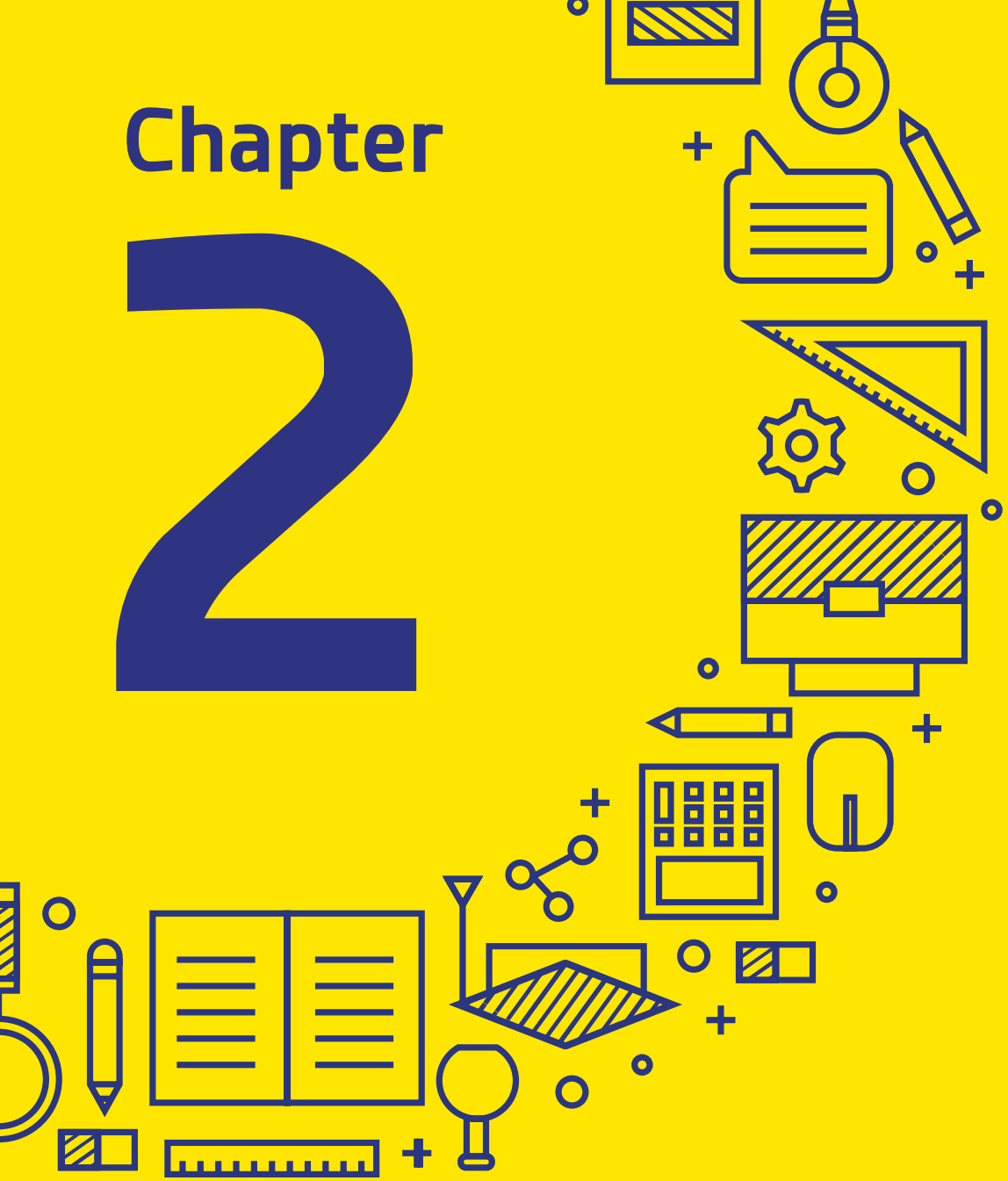
Finally, chapter 7 provides an overall discussion of the findings of chapters 2 to 6 and concludes with implications for theory and practice.

Note: A version of Chapter 2 was published as

Boevé, A. J., Meijer, R. R., Albers, C. J., Beetsma, Y., & Bosker, R. J. (2015). Introducing computer-based testing in high-stakes exams in higher education: Results of a field experiment. *PloS one, 10*(12), doi:10.1371/journal.pone.0143616

Chapter

2



Implementing Computer-Based Exams in Higher Education: Results of a Field Experiment

2.1 Introduction

Computer-based exams (CBE) have a number of important advantages compared to traditional paper-based exams (PBE) such as efficiency, immediate scoring and feedback in the case of multiple-choice question exams. Furthermore CBE allow more innovative and authentic assessments due to more advanced technological capacities (Cantillon, Irish, & Sales, 2004; Csapo, Ainley, Bennett, Latour, & Law, 2012). Examples are the use of video clips and slide shows to assess medical students in surgery (El Shallaly, & Mekki, 2012) or the use of computer-based case simulations to assess social skills (Lievens, 2013). However, there are also drawbacks when administering CBE such as the additional need for adequate facilities, test-security, back-up procedures in case of technological failure, and time for staff and students to get acquainted with new technology (Cantillon, Irish & Sales, 2004). Nevertheless, there have been concerns about equality of test-modes, fairness, and the stress students might experience (Whitelock, 2009)

In order to ensure a smooth transition to computer-based examining in higher education, it is important that students perform equally well on computer-based and paper-based administered exams. If, for example, computer-based administration would result in consistently lower scores than paper-based administration, due to unfamiliarity with the test mode or due to technical problems this would result in biased measurement. Thus, it is important that sources of error, or construct irrelevant variance (Huff & Sireci, 2001), which may occur as a result of administration mode, are prevented or minimized as much as possible in high-stakes exams. As will be discussed below, however, it is unclear from the existing literature whether the different administration modes will result in similar results.

The adoption and integration of computer-based testing in higher education has progressed rather slowly (Deutsch, Herrmann, Frese & Sandholzer, 2012). Besides institutional and organizational barriers, an important implementation consideration is also the acceptance of CBE by the students (Deutsch et al, 2012; Terzis & Economides, 2011). However, as Deutsch et al. (2012) discussed "little is known about how attitudes toward computer based assessment change by participating in such an assessment". Deutsch et al (2012) found a positive change in students' attitudes after a computer-based assessment. As with many studies in prior research (e.g., Deutsch et al., 2012; Terzis & Economides, 2011), this took place in the context of a mock exam that was administered on a voluntary basis. There is little research on student attitudes in the context of high-stakes exams, where students do not take the exam on a voluntary basis.

The aim of the present study was to extend the literature on high-stakes computer-based exam implementation by (1) comparing student performance on CBE with performance on PBE and (2) evaluating students' acceptance of computer-based exams. Before we discuss the design of the present study, however, we first discuss prior research on student performance, and acceptance of computer-based multiple-choice exams. The present study is limited to multiple-choice exams as using computer-based exams in combination with open-question or other format tests, may have different advantages or disadvantages, and the aim of this paper was not to study the validity of various response formats.

2.2 Student performance in computer and paper-based tests

The extent to which different administration modes lead to similar performance in educational tests has been investigated for different levels of education. A meta-analysis on test-administration

mode in K-12 (primary and secondary education) reading education showed that there was no difference in performance between computer-based and paper-based tests (Wang, Jiao, Young, Brooks, & Olson, 2008). A meta-analysis on computer-based and paper-based cognitive test performance in the general population (adults) showed that cognitive ability tests were found to be equivalent in different modes, but that there was a difference in performance on speeded cognitive processing tests, in favor of paper-based tests (Mead & Drasgow, 1993). In the field of higher education, however, as far as we know meta-analyses have not been conducted and results from individual studies seem to vary.

To illustrate the diversity of studies conducted, Table 2.1 shows some characteristics of a number of studies investigating difference in performance between computer-based and paper-based tests with multiple-choice questions in the context of higher education. The studies vary in the number of multiple-choice questions included in the exam, in the extent to which the exam was high-stakes, and in the extent to which a difference in performance was found in favor of a computer-based or paper-based mode of examining. While our aim was not to conduct a meta-analysis, Table 2.1 also shows that many studies do not provide enough statistical information to compute an effect-size. Furthermore, not all studies include a randomized design, implying that a difference cannot be causally attributed to mode of examining. Given these varying findings, establishing that administration mode leads to similar performance remains an important issue to investigate.

Table 2.1. Studies investigating performance differences between paper-based and computer-based tests with multiple-choice questions

	Number of mc-questions	Randomized	High-stakes	Effect size (Cohen's *d*)	Result in favor of
Lee & Weekaron, (2001)	40	no	yes	0.69	paper-based
Clariana & Wallace (2002)	100	yes	yes[a]	0.76	computer-based
Cagiltay & Ozalp-Yaman (2013)	20	yes	yes	0.15	computer-based
Bayazit & Askar (2012)	6	yes	unclear	0.32	paper-based
Nikou & Economides (2013)	30	yes	unclear	0.19	computer-based
Anakwe (2008)	25	no	yes	not possible	
Frein (2011)	3	no	unclear	not possible	
Ricketts & Wilks (2002)	unclear	no	yes	not possible	
Kalogeropoulos et al. (2013)	unclear[b]	yes	unclear	not possible	

[a]the test counted for 15% of the final grade
[b]5 mc-items - but reported means for the mc-test are larger than 5

2.3 Student acceptance of computer-based tests

It is important to understand student acceptance of computer-based testing because the test-taking experience is substantially different from paper-based exams (McDonald, 2002). In paper-based exams with multiple-choice questions, several questions are usually presented per page, and students have the complete exam at their disposal throughout the time allotted to complete the exam. Common test-taking strategies for multiple-choice exams include making notes, marking key words in specific questions, and eliminating answer categories (Towns & Robinson, 1993; Kim & Goetz, 1993). In computer-based multiple-choice exams, however, standard software may not offer these functionalities. For an example where these functionalities were excellently included see McNulty et al. (2007).

Apostolou, Bleu, and Daigle (2009) found mostly negative appraisals of computer-based testing by students in accounting, recommending more research be conducted into what aspects of computer-based exams are important to students. In a mock-exam environment, Wibowo, Grandhi, Chugh, and Sawir (2016) found that most students experienced the exam in a computer-based mode as more stressful compared to the paper-and-pencil mode of examining. While about three quarters of the students who participated in this study were willing to take a digital exam in the future, about half of the students clearly still preferred a paper-based exam. Dermo (2009) investigated student perceptions of the computer-based mode of examining, including both the formative and summative contexts, and found that on average students opinions were rather neutral towards mode of examining. While students were not invited to clearly indicate whether they preferred a particular mode of examining, qualitative comments gave the impression that students in the Dermo (2009) study preferred the CBT mode. A limitation of prior research is that evaluation of computer-based tests has sometimes been confounded with the evaluation of other aspects of testing not directly related to the computer-based testing mode. In the studies of Peterson and Reider (2002), as well as Dermo (2009), the operationalization of student perceptions implies that using computer-based testing means increased testing with multiple choice questions rather than open questions. As a result, the outcomes of these studies may reflect students opinions concerning multiple-choice versus open questions rather than their perceptions of examination mode.

A study by Hochlehnert et al. (2011) in the German higher education context showed that only 37% of students voluntarily chose to take a high-stakes exam via the computer, and that test-taking strategies were a reason why students opted for the paper-based exam. Deutsch et al. (2012) showed that the attitudes of medical students in Germany became more positive towards computer-based assessment after taking a practice exam. The context in which students take a mock exam however, is very different to the actual environment of a formal high-stakes exam. Therefore, it is important to investigate both the test-taking experience and student acceptance of computer-based exams in a high-stakes exam.

Based on focus-group interviews, Escudier, Newton, Cox, Reynolds, and Odell (2011) found that students experienced both advantages and disadvantages in making computer-based multiple choice tests. Among the advantages were, for example, the ease of changing answers, and the prevention of cheating. Advantages of the paper-based mode were, for example, the overview over the whole exam, and making notes and highlighting

in the questions. Although most students found the digital assessment acceptable, almost 25% thought it was not acceptable or less optimal than other methods, and 10% of students thought the computer-based mode was unfair.

As one of the few studies in the context of high-stakes exams, Ling, Ong, Wilder-Smith and Seet (2006) found that students preferred the computer-based mode of examining particularly for multiple-choice exams, and less so for open-question exams. Student response rates however were rather low in this study since students were contacted by e-mail after the exam, which means the results could have been biased if students who did not have a mode preference or a paper-and-pencil preference were less likely to respond to the questionnaire.

2.4 The present study

The present study took place in the last semester of the academic year 2013/2014 with psychology students in the first year of the Bachelor in Psychology program (Dutch track), and was replicated in the academic year of 2014/2015 with a new cohort of students following the International track.

The university opened an exam facility in 2012 to allow proctored high-stakes exams to be administered via the computer. In the academic year 2012/2013 there were 101 computer-based exams, and this number increased to 225 exams in 2013/2014. Of these exams, 102 were multiple-choice exams, 155 were essay question exams, 58 were a mix of both formats, and 11 exams were in a different format. Most computer-based exams were implemented via the university's online learning platform NESTOR, which is embedded in Blackboard (www.blackboard.com), but has extra programming modules developed by the university. Within the broad project to implement computer-based exams, an additional collaboration of faculties started a pilot project to facilitate computer-based exams through the Questionmark Perception (QMP) software (www.questionmark.com). Of the multiple-choice exams administered over the two-year period, 62 were administered via QMP and 40 were administered via Blackboard. Nevertheless, the program of psychology had no previous experience with computer-based examining.

The psychology program is a face-to-face based program (in contrast to distance learning). However, for the course that was included in the present study, attending lectures was not mandatory, and students had the option to complete the course based on self-study alone, given that they showed up for the midterm and final exam.

2.5 Method

To evaluate student performance in different exam modes and acceptance of computer-based exams, computer-based examining was implemented in a biopsychology course, which is part of the undergraduate psychology program. Assessment of the Biopsychology course consisted of two exams receiving equal weight in grading; both were high-stakes proctored exams. Since the computer-based exam facilities could not facilitate the whole group of students, half of the students were randomly assigned to make the midterm exam by computer, and the other half of the students were assigned to make the final exam by computer.

In order to examine whether there were mode differences in student performance

on both exams, we analyzed student performance. Student performance data was collected by the University of Groningen for academic purposes. In line with the university's privacy policy, these data can be used for scientific research when no registered identifiable information will be presented. Since the analysis of student grades presented in this study entailed comparing summary measures of student grades for particular exam mode, no registered identifiable information was presented. Therefore, written informed consent for the use of student grades for scientific research purposes was not obtained.

In order to examine student acceptance of computer-based exams, a questionnaire was placed on the exam desks of students, which they could voluntarily fill out, with the knowledge that their response to the evaluation questionnaire could be used for scientific purposes. Furthermore, students were notified of this procedure at the onset of the course. We did not ask students for written informed consent as to whether they were willing to fill out the questionnaire since they were able to choose to fill out the questionnaire voluntarily and anonymously. Since students were aware that their responses would be used for scientific purposes, informed consent was implied when students chose to fill out the questionnaire. This study, including the procedure for informed consent, was approved by, and adhered to the rules of the Ethical Committee Psychology of the University of Groningen[1].

In the psychology program, this was the first time a computer-based exam was implemented. The total assessment of the course in biopsychology (in both years the study was conducted) consisted of a midterm and final exam, which both contributed equally to the final grade. These exams took place in a proctored exam hall. At the start of the course students were randomly assigned to make the midterm exam either by computer or as a paper-and-pencil test. Subsequently the mode of examining was switched for the final exam, so that everyone was assigned to take either the midterm or the final exam as a computer-based test. After completing the computer-based exam, students were invited to fill-out a paper-and-pencil questionnaire on their experience with the computer-based exam, which they could submit before leaving the exam hall. Students received immediate feedback on their performance on the exam in the computer-based condition (number of questions correct), and thus knew their performance on the exam when completing the questionnaire. Students in the paper-based condition received the exam result within a couple of days after taking the exam.

2.5.1 Participants

At the start of the course in the 2013/2014 study there were 401 students enrolled via the digital learning environment. These students were randomly assigned to make the midterm exam via paper-based mode or computer-based mode. If a student was assigned to complete the midterm on paper, the final exam would be completed by computer and vice versa. In the 2014/2015 study there were 428 students enrolled in the course, and these students were also randomly assigned to take the midterm via paper-based mode or computer based mode as in the 2013/2014 study. All students who completed a computer-based exam were

[1] http://www.rug.nl/research/heymans-institute/organization/ecp/

invited to evaluate their experience by responding to a paper-based questionnaire directly after completing the CBE. As can be expected in a field experiment, however, there was both some attrition and non-compliance (Figures 2.1a and 2.1b), which we will discuss below.

There were three sources of attrition in the first study (Dutch cohort 2013/2014): (1) not registering for the exams, (2) registering but not showing up at the midterm, and (3) completing the midterm but not showing up for the final exam. These three sources of attrition led to a 16% overall attrition rate (66 students). There were 16 students who completed both the midterm and final exam on paper. In addition, there was a technical failure at the midterm exam, as a result of which 36 students needed to switch to a paper-based exam in order to be able to complete the exam.

In the international cohort of 2014/2015, there were also 3 sources of attrition: students who did not show up for either exam, students who did not show up to the midterm exam, and students who did not show up to the final exam. Note that in 2013/2014 students were required to both enroll in the course and register for the exam separately, while in 2014/2015 the system had changed so that course enrollment automatically implied exam registration. The overall attrition rate in the 2014/2015 cohort was 10% (44 students). There was no technical failure, and one student completed both midterm and final exam in the paper-based mode.

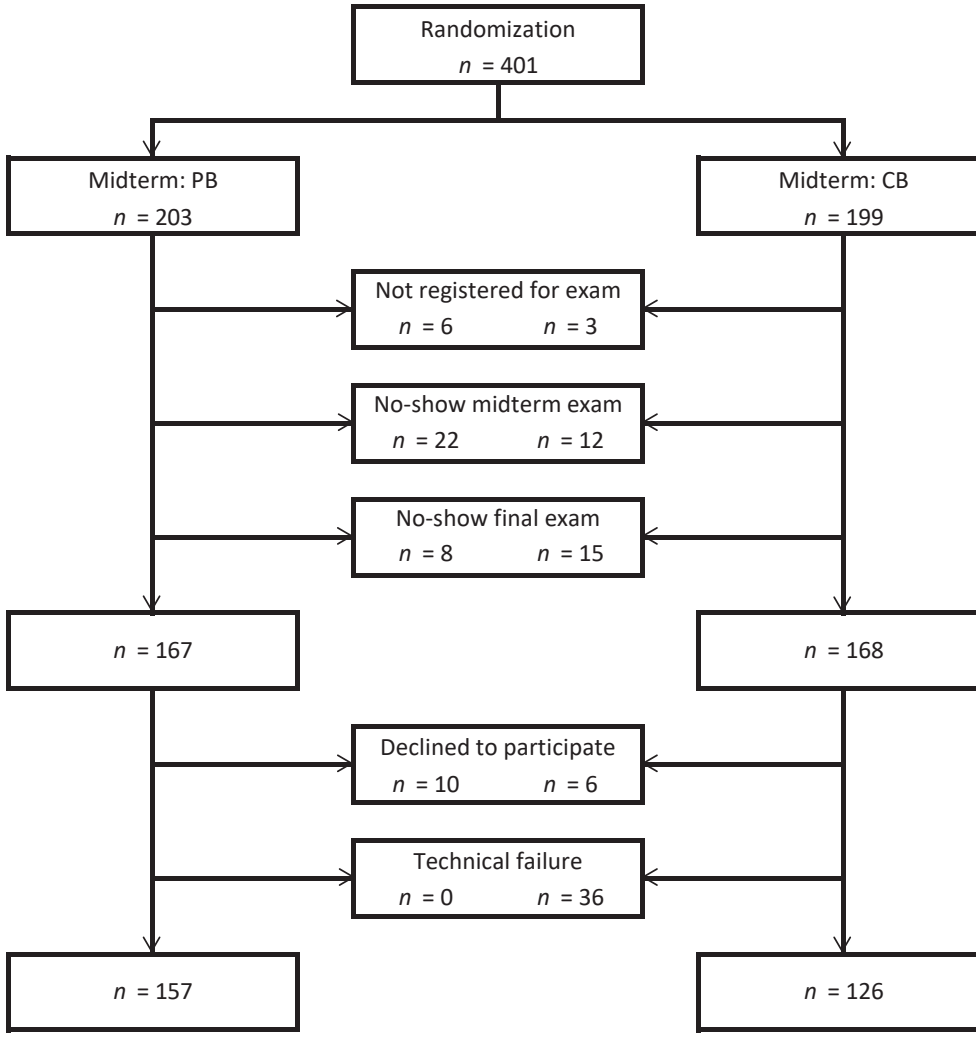


Figure 2.1a. Random assignment of students to different exam modes, subsequent attrition, and non-compliance for the first study (Dutch cohort 2013/2014)

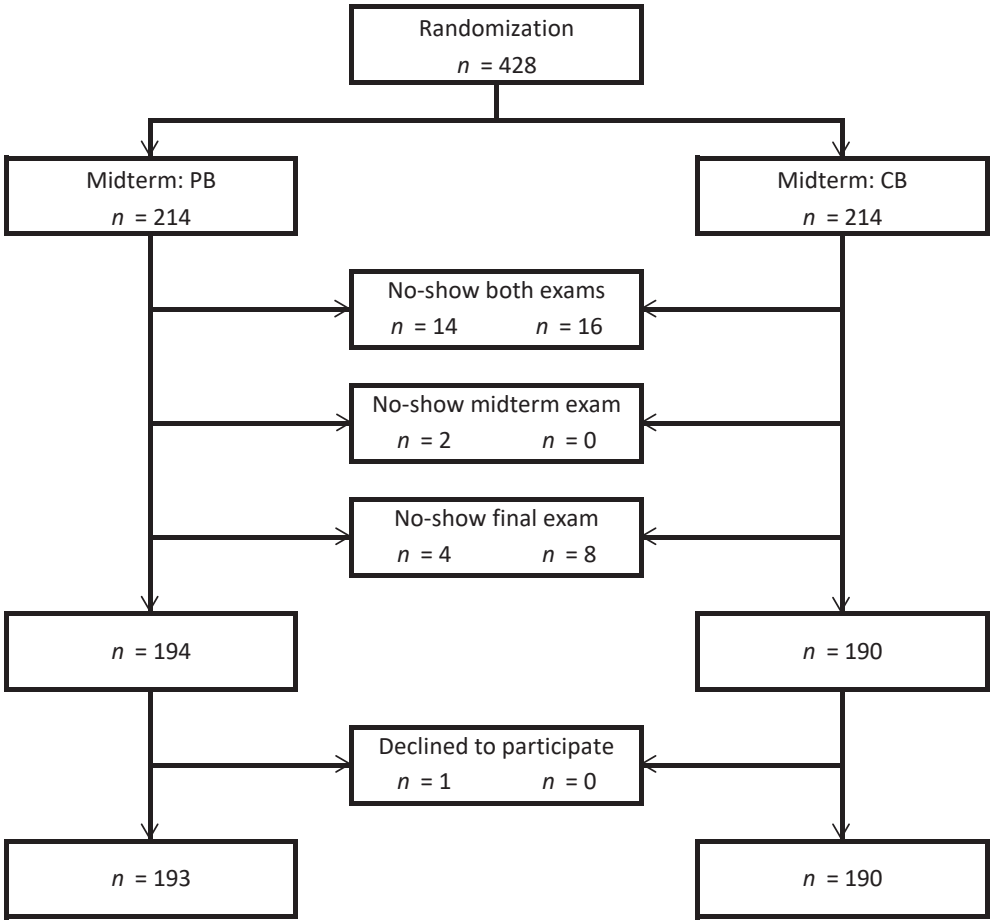


Figure 2.1b. Random assignment of students to different exam modes, subsequent attrition, and non-compliance for the second study (international cohort 2014/2015)

2.5.2 Materials

Student performance. Both the midterm and final exam contained 40 multiple-choice questions with four answer categories. The exams measured knowledge of different topics in biopsychology. The material that was tested on the midterm exam, was not tested again in the final exam. Thus the two exams covered different material included in the course and each exam had an equal weight in determining the final grade. The midterm exam appeared to be somewhat more difficult (mean item proportion correct) in both cohorts (see Table 2.2). Student performance in both modes was investigated by comparing the mean number of questions correct on each exam.

Table 2.2. Exam characteristics of both cohorts, and partial exams, with mean item-total correlations and reliability estimates for the computer-based (CB) and paper-based (PB) mode.

	Proportion correct (Mean)	**Item-total correlation (Mean)**		**Reliability [95% CI]**	
2013/2014		CB	PB	CB	PB
Midterm	.72	.32	.29	.78 [.72; .83]	.71 [.66; .76]
Final	.75	.32	.27	.75 [.67; .80]	.66 [.59; .73]
2014/2015					
Midterm	.77	.33	.31	.82 [.79; .86]	.80 [.76; .84]
Final	.80	.33	.35	.82 [.79; .86]	.84 [.81; .87]

Acceptance of computer-based tests. Student acceptance was operationalized in three ways (see Table 2.3). First, students answered questions about their test-taking experience during the computer-based exam and in paper-based exams in general. Second, students were asked whether they preferred a computer-based exam, paper-based exam, or did not have a preference. Third, students were asked whether they changed their opinion about computer-based exams as a result of taking a computer-based exam. Answers to the questions on test-taking experience were given on a five-point Likert response scale ranging from ‘completely disagree’ to ‘completely agree’. The question on whether students’ opinions changed had five response options: ‘yes, more positive’, ‘yes, more negative’, ‘no, still positive’, ‘no, still negative’, and ‘no, still indifferent’.

Table 2.3. Evaluations of students test-taking experience and acceptance of computer-based exams

Student acceptance of computer-based exams	
Questions	**Sub-questions**
In this computer-based exam	I was able to work in a structured manner
	I had a good overview of my progress in the exam
	I was able to concentrate well
In paper-based exams in general	I am able to work in a structured manner
	I have a good overview of my progress in the exam
	I am able to concentrate well
I prefer a: paper-based exam, computer-based exam, no preference	
Did your opinion about computer-based ex-ams change as a result of taking this exam?	

2.5.3 Procedure

The midterm computer-based exam in the 2013/2014 cohort was administered through the Questionmark software, but as mentioned above, there was a technical problem. Since the technical issue could not be solved in time, the final exam was administered directly via Nestor (the university's online learning platform). As a result of the change in interface, the design and layout of the computer-based midterm and final exam was slightly different in the 2013/2014 cohort. The midterm exam, administered through QMP, was designed so that all questions were presented simultaneously with a scrolling bar for navigation. In the final exam, administered via Nestor, the questions were presented one at a time and navigation through the exam was done via a separate window with question numbers allowing students to review and change answers given to other questions. In the 2014/2015 cohort, both the midterm and final computer-based exam was directly administered via Nestor, in the same way as the final exam in the 2013/2014 cohort.

For both partial exams in both cohorts students had the opportunity to go back and change their answers at any point and as many times as they liked before submitting their final result. After submitting their final answers to both the midterm and final exam in the computer-based mode, students immediately received an indication of how many questions they answered correctly. For the paper-based mode of examining, students took a list of their recorded answers home, and could calculate an indication of how many questions they answered correctly several days after the exam when the answer key was made available in the digital learning environment.

2.6 Results

2.6.1 Student Performance

Table 2.4 shows that there was no statistically significant difference in the mean-number of questions answered correctly between the computer-based and paper-based mode for both the midterm and final exam in both 2014 and 2015.

Table 2.4. Mean number of questions correct in the different exam conditions for the midterm and final exam

	Computer-based		**Paper-based**			
2013/2014	*n*	*M*(*SD*)	*n*	*M*(*SD*)	*t*(*df*)	Cohen's *d* [95% CI]
Midterm	126	28.56 (5.31)	157	28.50 (4.61)	0.10 (281)	0.01 [-0.22; 0.25]
Final	157	29.92 (4.60)	126	29.50 (4.32)	0.78 (281)	0.09 [-0.14; 0.33]
2014/2015						
Midterm	190	31.11 (5.69)	193	30.82 (5.47)	0.50 (381)	0.05 [-0.15; 0.25]
Final	193	32.12 (5.35)	190	32.05 (5.72)	0.13 (381)	0.01 [-0.19; 0.21]

2.6.2 Student acceptance of CBE

Test-taking experience. In Figure 2.2 the mean scores on the questions with respect to test taking experiences for the midterm and final exam are provided. A multivariate ANOVA was conducted to examine whether these questions were evaluated differently for the midterm and final exam. Results of the overall model test ($\alpha = .05$) showed that there was a difference in how the questions were evaluated between the midterm and final exam (2013/2014: $F(6,258) = 7.02$, $p < .001$, partial-$\eta^2 = .14$; 2014/2015: $F(6,320) = 3.87$, $p = .001$, partial-$\eta^2 = .07$).

Additional (Bonferroni corrected α=.0083) univariate analyses showed that students in the 2013/2014 cohort were less able to concentrate in the midterm computer-based exam compared to the final computer-based exam ($F(1,265) = 22.39$, $p = .00014$). Students in the 2014/2015 cohort on the other hand were less able to monitor their progress ($F(1,325) = 11.78$, $p = .0007$) and concentrate ($F(1,325) = 11.39$, $p = .0008$) in the computer-based final exam compared to the computer-based midterm exam. See Table A2 in the appendix for more details on the means, standard-deviations, and effect sizes.

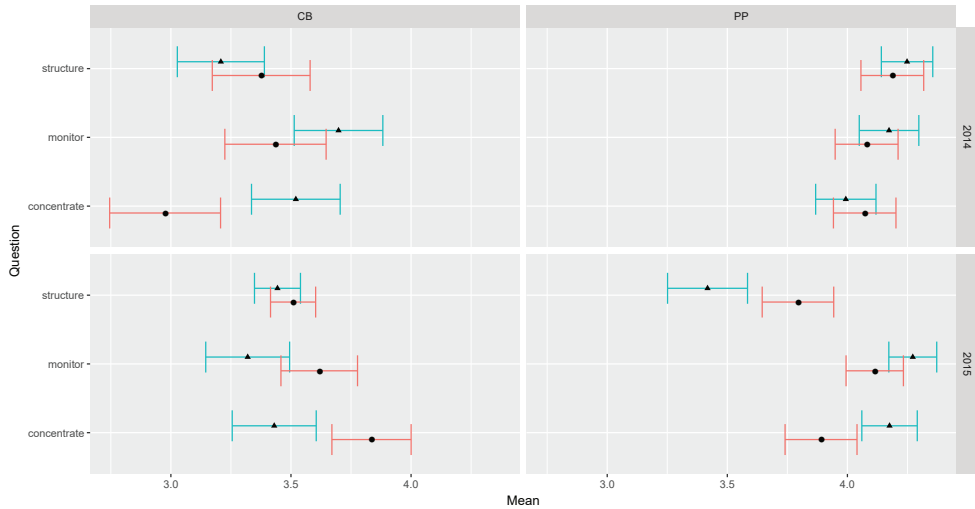


Figure 2.2. Mean scores and 95% confidence intervals for student approaches to completing the computer-based exam, and paper-based exams in general for the midterm (red line) and final exam (blue line).

To examine the difference in test-taking experience between the computer-based exam and paper-based exams in general, Bonferroni corrected (α = .017) dependent-sample *t*-tests were conducted. Table 2.5 shows that students were more positive in terms of their ability to work in a structured manner, monitor their progress, and concentrate during paper-based exams compared to the computer-based exam, with medium (0.33) to large (0.64) effect sizes.

Table 2.5. Mean difference between computer-based and paper-based exam evaluation, with dependent-sample *t*-test results and effect-size

2013/2014	**CB – PB mode [95% CI]**	***t*(*df*)**	**Cohen's *d* [95% CI]**
Structured approach	-0.94 [-1.12; -0.77]	-10.71 (268)	-0.65 [-0.83; -0.48]
Monitor progress	-0.54 [-0.72; -0.37]	- 6.15 (269)	-0.37 [-0.54; -0.20]
Concentration	-0.74 [-0.92; -0.56]	- 8.13 (269)	-0.46 [-0.67; -0.32]
2014/2015			
Structured approach	-0.72 [-0.87; -0.57]	-9.64 (333)	-0.53 [-0.68; -0.37]
Monitor progress	-0.40 [-0.57; -0.23]	-4.75 (334)	-0.26 [-0.41; -0.11]
Concentration	-0.43 [-0.57; -0.30]	-6.22 (332)	-0.34 [-0.49; -0.19]

Preference for computer-based exams. In the 2013/2014 cohort, 50% of the students preferred a paper-based exam, 28% preferred a computer-based exam, and 22% indicated that they did not have a preference for one mode over another after completing the computer-based exam. There was no difference in preference for a particular exam-mode between students who completed the midterm and final exam via the computer

(Fisher's exact p = .97). In the 2014/2015 cohort, there was a difference in preference for exam mode between students who completed the midterm and final exam via the computer (Fisher's exact p =.007). For students who took the midterm by computer, 38.5% preferred a computer-based exam, 38.5% preferred a paper-based exam, and 23% did not have a preference. For students who took the final exam by computer, 23% preferred a computer-based exam, 51% preferred a paper-based exam, and 26% did not have a preference.

With respect to the change of opinion towards computer-based assessment after taking a computer-based exam, in the 2013/2014 cohort: 43% of students felt more positive, 14% felt more negative, 15% were indifferent, 16% still positive, and 12% still negative towards computer-based exams, with no difference between the midterm and final exam (Fisher's exact p = .12). In the 2014/2015 cohort, of the group who took the midterm by computer 44% were more positive, 9% more negative, 16% indifferent, 24% still positive, and 6% still negative towards taking computer based exams. Of the group who took the final exam by computer however, 44% were more positive, 22% more negative, 11% indifferent, 13% still positive, and 11% still negative towards taking computer-based exams consisting of multiple-choice questions.

2.7 Discussion

2.7.1 Student performance

In line with recent research (Cagiltay & Ozalp-Yaman, 2013; Bayazit & Askar, 2012; Nikou & Economides, 2013), we found no difference in the mean number of questions correct between computer- and paper-based tests for both the midterm and final exam. Earlier findings in the field of higher education in favor of paper-based tests (Lee & Weekaron, 2001), and in favor of computer-based tests (Clariana & Wallace, 2002), were not replicated in this study. Based on these findings, we can conclude that recent findings show that exam-mode may not cause differential student performance in higher education. An important explanation for this finding could be the population of students in this study. Students in this study entered the higher education system largely directly after completing secondary education and represent a generation that has grown up with technology. Earlier studies on the use of computer-based testing may have found a difference in favor of paper-based tests as a result of test takers' unfamiliarity with technology. Therefore, the lack of a difference in performance between modes in the present study may be the result of a generational difference in student population compared to older studies. This also implies that current studies with older populations of students may still find a mode effect, although adults today will have had more technology exposure in daily life than studies conducted with adults twenty years ago.

2.7.2 Student acceptance of CBE

Students indicated that the test-taking experience in PBE in general was more favorable compared to CBE in terms of their ability to work in a structured manner, have a good overview of their progress through the exam, and their ability to concentrate. While there was no difference in performance for computer-based and paper-based exams, these findings suggest that students appear to feel less in control when taking a computer-based exam relative to a paper-based exam. This is in line with previous findings by Hochlehnert et al. (2011) who found that the

absence of functionality to apply test-taking strategies was a reason for students not to choose a computer-based exam. Further research is necessary to see if this difference in approach to taking the exam may be an artefact of the first-time introduction to computer-based exams. Students who regularly take computer-based exams may be more accustomed to this mode, and therefore have developed confidence in their approach to taking computer-based exams. Another avenue that may be pursued in order to better understand the test-taking experience in CBE may be to extend the research of Noyes, Garland and Robbins (2004) who found that students experienced a higher cognitive load in a short computer-based multiple-choice test compared to an equivalent paper-based test. Further research could investigate the extent to which the perceived test-taking experience is related to cognitive load.

In the 2013/2014 cohort, we found that students who took the final exam by computer, were able to concentrate better on average than students who took the midterm exam by computer. A possible explanation for this result may be the technical problem during the midterm. Students in the computer-based exam hall who did not experience the technical problem, may have been affected indirectly by the unrest in the exam hall as the directly affected students were provided with a paper-based exam. If this were the explanation for the difference in concentration between the midterm and final exam, it would seem logical that students who completed the midterm exam were also more negative about computer-based exams compared to the group of students who completed the final exam by computer. We found no difference, however, in the extent to which student opinions became more negative towards CBE after taking the computer-based exam.

Another possible explanation for the difference in the ability to concentrate between the midterm and final exam in the 2013/2014 cohort is the design of the computer-based assessment. According to Ricketts and Wilks (2002) a difference in design from scrolling through all questions to a one-question-at-a-time format explained improved student performance. In the present study all the questions were displayed simultaneously in the midterm file, while in the final exam questions were presented one at a time. In presenting questions one at a time during the final exam students may have been able to focus better on the questions at hand, explaining the greater ability to concentrate reported by students.

The replication in 2014/2015 showed a larger difference in experience between students who took the midterm exam by computer and students who took the final exam by computer, with students who took the midterm exam by computer generally being more favorable about their experience in the CBT compared to students who took the final exam by computer. Furthermore, this difference was also reflected in a difference in student preference for exam mode between students who took the midterm or final by computer. In terms of ability to concentrate in the computer-based exam, results were different from the 2013/2014 findings, namely students in the midterm computer-based exam indicated being able to concentrate better than students in the final computer-based exam. In the 2014/2015 cohort, however, both the midterm and final exam were designed to present one question at a time, and thus no difference in exam experience would be expected. There were no technical failures, or problems with the administration of computer-based exams in 2014/2015, therefore the difference in experience between students who took the midterm and final by computer is difficult to explain.

2.7.3 Practical implications

Based on the above discussion there are several practical implications for Universities seeking to implement CBE. Student performance on multiple choice question exams does not appear to vary across test mode. The benefits of CBE, and the lack of negative consequences, can both be used in the communication towards students prior to the first implementation of CBE in order to maximize acceptance. Furthermore, universities need to invest in good CBE exam facilities. This includes investing in adding more test-taking functionalities so that students test-taking experience may be as optimal as possible. Furthermore, the potential of technical failure is a risk that requires good protocols so that students are able to complete the exam either on a different computer or on paper.

The full potential of computer-based tests can be realized in further developments. One option is to use computer adaptive testing (CAT). The advantage of CAT is that items are chosen from an item pool that best fit the level of the candidate. In many higher education institutes however, this is difficult to realize as a very large item pool with regular refreshment is needed. In combination with the extensive psychometric knowledge necessary for this development, this is generally beyond the scope of many university courses. What may be easier to realize however, is to offer test items to students in random order, which helps prevent cheating.

2.7.4 Limitations

There were several limitations to the present study. First, there were technical problems during the midterm computer-based exam in the 2013/2014 cohort. As a result of this technical failure a number of students had to complete the planned CBE on paper. This remains a risk for computer-based exams in general, and the facilities for computer-based examining need to be organized in such a way that when this occurs unexpectedly in practice, hindrance for students is minimized. In the present study, students were allowed extra time to complete the exam, although no one made use of it. It is important to note, that while students may not have a good test-taking experience, their results are unlikely to suffer as a consequence. Several studies have shown student performance in CBE is not affected by technical issues (Sinharay, Wan, Choi & Kim, 2015; Sinharay, Wan, Whitaker, Kim, Zhang, & Choi, 2014).

An important aspect of introducing computer-based assessment deserves mention as well, namely the teacher of faculty perspective. Since the present study was conducted in a single course, the teacher perspective was outside the scope of the present study. Research into teacher acceptance and willingness to implement computer-based assessment may also provide relevant insight into improving the implementation of computer-based exams in higher education.

Furthermore, our sample consisted of students who were primarily ‘traditional’ students and started their study soon after completing high school. A population containing more mature aged students may view technology differently. In addition students were studying face to face. Students who study via distance mode may view computer-based testing differently than face-to-face students.

2.7.5 Conclusion

This study found that students performed equally well in computer-based multiple-choice exams compared to paper-based exams. While paper-based exams may be the norm in many universities, investing in computer-based exams may be beneficial for the younger generation who are more and more growing up with computer and digital technologies. Further research is necessary into the optimal design of computer-based exams, such that student-acceptance is maximized and not an irrelevant source of stress during exams in a high-stakes context.

Note: This chapter was published as

Meijer, R. R., Boevé, A. J., Tendeiro, J. N., Bosker, R. J., & Albers, C. J. (2017). The Use of Subscores in Higher Education: When Is This Useful? *Frontiers in Psychology*, *8*, 305. doi:10.3389/fpsyg.2017.00305.

Chapter

3

Using Subscores in Higher Education

3.1 Introduction

For teachers in higher education, student assessment through administering and scoring exams is a common and efficient method to test large groups of students. Cizek (2009) defined a test (or exam) as "a systematic sample of a person's knowledge, skill, or ability" and assessment as a much broader planned process of gathering such information for different purposes. Assessment in higher education is challenging for teachers since they face more students, with less contact-time compared to teachers in primary and secondary education. Using a single test for multiple purposes in assessment is, therefore, an efficient way of assessment. Providing students with feedback is often suggested to improve the quality of learning, and thereby increasing student performance (Black & William, 1998; 2003). One way to provide feedback while keeping teacher burden low, is to report subscores, that is, to report the sum of item scores on a specified number of items, because it is assumed that these subscores may provide additional information to the total score on the exam. This idea is not new and there are many examples where subscores on exams or tests are used for diagnostic, formative, and remedial purposes (e.g., Harks, Klieme, Hartig, & Leiss, 2014; Ketterlin-Geller & Yovanoff, 2013; Schneider & Andrade, 2013). For example, the total score on a reading comprehension test may be reported together with subscores that reflect specific reading skills, like being able to understand the meaning of a story as opposed to being able to read and understand individual sentences (Reckase & Xu, 2014).

In large-scale testing, reporting subscores is sometimes even required. For example, in the US for some educational programs it is required that "students should receive diagnostic reports that allow teachers to address their specific academic need; subscores could be used in such a diagnostic report" (Sinharay et al., 2010, p. 150). In primary education in the Netherlands the use of subscores for different topics like reading comprehension and arithmetic is required for the general test that helps to determine which type of secondary education students will follow (Rijksoverheid, 2015).

Before reporting subscores, teachers and instructors should provide evidence that observed subscores contain unique information over and above the observed total score in terms of the true subscores. In the often cited Standard 1.14 of the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014) it is said that "When interpretation of subscores, score differences, or profiles is suggested, the rationale and relevant evidence in support of such interpretation should be provided" and "When a test provides more than one score, the distinctiveness and reliability of the separate scores should be demonstrated, and the interrelationship of those scores should be shown to be consistent with the construct(s) being measured" (p. 27). Like incorrect or invalid test scores may have serious detrimental effects on grading or selection, unreliable and invalid subscores may have detrimental effects on decisions made to assign students to remedial teaching groups or to invest more time in particular knowledge domains.

For commercial tests and questionnaires, techniques like factor analysis and scale analyses are often used to investigate whether it is useful to distinguish separate clusters of item scores. Because users of large-scale tests are expected to justify the interpretation of subscores, the relevance of investigating the quality of subscores is clear in this context. It

is, therefore, also not surprising that recent psychometric studies in large-scale educational testing (e.g., Sinharay, Puhan & Haberman, 2011) discussed when subscores provide additional information to the total scores. However, many of these studies are rather technical and aimed at educational researchers. As a result, these papers are often difficult to understand for practitioners.

As Cizek (2009) argued, however, the context of classroom assessment is different from the context of large-scale assessment. The rigorous and extensive test-development techniques of large-scale tests are not generally used for classroom tests. An important reason for the latter is that, in general, stakes are lower in classroom testing than in large-scale testing. In higher education, however, tests results sometimes determine whether a student can follow another course or will suffer financial consequences from study delay. If any information on item and test quality is given to or monitored by teachers in higher education, this is by through reporting classical indices like proportion-correct scores, item-total correlations, and reliability estimates. As we will demonstrate below, the classical test-theory framework can also be used to evaluate the quality of subscores in classroom tests in higher education.

In this chapter we analyzed two exams from a psychology program with a method that can be used to investigate whether subscores have added value over and above total scores. Using this method may help to judge whether it is useful to report subscores for different tests used in higher education. We used both a multiple choice exam and an exam that consisted both of multiple choice items and open-ended questions. For this latter exam we also investigated the added value of the open-ended questions to the multiple choice questions in terms of measurement precision.

This chapter has the following structure. We first discuss an existing method that can be used to investigate whether subscores have added value. Second, we analyzed the college exams. Finally, we discuss the implications of our study for formative assessment in psychology education. In this paper we use psychometric arguments; every teacher or instructor is, of course, free to decide that information obtained from subscores is still useful irrespective of the outcome of a psychometric analysis. However, we think that it may be illuminating to see that information obtained from subscores that seems intuitively useful may not contain additional information over and above the total score.

3.1.1 Rationale behind the added value of subscores

We used a method discussed by Haberman (2008). Assume that we have an exam and that we calculate the total score on this exam as the number of questions answered correctly. Furthermore, assume that we are interested in reporting subscores on subsets of items. Haberman's method (to be discussed in more detail below) is based on two important psychometric indicators to determine whether or not subscores may have added value to the total score. The first is the correlation between the (true) subscore and the total score and the second is the estimated reliability of the total score and the individual subscores. The idea is that when the reliability of the individual subscores is relatively low, often due to a limited number of items, and the correlation between subscores and the total score is relatively high, reporting subscores in addition to reporting the total score has no *added* value over and above reporting only the total score.

Sinharay (2010) reviewed a number of large-scale exams administered in the US and concluded that "subscores on operational tests have more often been found not to be useful than to be useful." He also noted that "there is a lack of studies that demonstrated the validity of inferences made from subscores." For example, there is lack of evidence that subscores are related to other external criteria and that the incremental validity of subscores is valuable when subtest scores are highly related. Based on empirical and simulation studies Sinharay (2010) concluded that:

(a) Subscores based on tests smaller than 10 items almost never have added value because the reliability of these subscores is often too low, and that

(b) "The most important finding is that it is not easy to have subscores that have added value. Based on the results here, the subscores have to consist of at least about 20 items and have to be sufficiently distinct from each other to have any hope of having added value. Several practitioners believe that subscores consisting of a few items may have added value if they are sufficiently distinct from each other. However, the results in this study provide evidence that is contrary to that belief. Subscores with 10 items were not of any added value even for a realistically extreme (low) disattenuated correlation of .7."

However, these rules-of-thumb were predominantly obtained from large-scale exams and it is unclear whether these results can be generalized when investigating the added value for classroom tests.

3.1.2 Method proposed by Haberman (2008)

As discussed above, in the present study we concentrated on a method suggested by Haberman (2008) that is based on classical test theory. Many tests used in higher education are evaluated using classical test theory indices and so this method can be easily applied in this context. To determine whether subscores have added value over and above the total score, Haberman (2008) used the proportional reduction in mean squared error (PRMSE). The central idea is that one should only use a subtest score over a total score when it can be shown that the observed subtest score leads to a larger reduction in mean squared error in estimating the true subtest score than the observed total score. It can be shown that this is the case when the correlation between the true subtest score and the observed subtest score is larger than the correlation between the true subtest score and the observed total score (Haberman, 2008). The larger the PRMSE, the smaller the mean squared error to estimate the true subscore.

Let $PRMSE_s$ denote the PRMSE associated with the regression estimate of the true subscore on subtest *s* by means of the observed subscore on subtest *s*. Let $PRMSE_x$ denote the PRMSE associated with the regression estimate of the true subscore on subtest *s* by means of the observed total score on test *x*. Haberman (2008) showed that $PRMSE_s$ equals the estimated reliability of the observed subscore. The idea is that the observed subscore provides added value over and above the observed total score to estimate the true subscore when the observed reliability of the subtest score ($PRMSE_s$) is larger than $PRMSE_x$. In the context of typical performance testing in psychology, Reise, Bonifay, and Havilund (2014) give a step-by-step instruction on how to calculate the $PRMSE_x$.

3.2 Method

We investigated the added value of subscores on two exams from a degree program in psychology. The exams were taken by second year bachelor's degree students at an international degree program in psychology at a Dutch University, in the academic year 2014-2015. Exam records were collected primarily for educational purposes and these existing data could be used for research purposes in accordance with this university's privacy policy.

The first exam (34 items) was from a course on test theory taken by 319 students. We chose to split the exam into two subtests, namely 14 items that could be classified as factual knowledge and 20 items that reflected conceptual understanding of test construction and test use. These subtests were classified after test-construction and in a subjective manner by the authors of this study. In the faculty where this research took place, there was an interest in using Bloom's taxonomy (Krathwohl, 2002) to give students feedback on the depth of their understanding. This inspired investigating whether classifying a test used in practice into subtests based on different types of knowledge would lead to subtest scores that provided more information than the total score.

The second exam was from a statistics course and consisted of 5 short-answer/ partial credit open-ended questions and 20 multiple choice questions, where the final grade was computed based on 25% of the score on the open-ended questions and on 75% of the score on the multiple choice questions. The exam was administered to 350 examinees that followed the course in the English language. For the open-ended part of the exam, a grade between 1 and 10 was assigned.

There is a large body of literature that shows that, in general, administering multiple choice questions is a more efficient way of measuring knowledge than open-ended questions and that open-ended questions are not superior to multiple choice items in terms of reliability and validity (e.g., Hift, 2014). However, both teachers and students are sometimes in favor of open-ended questions. One of the main reasons for teachers to use open-ended questions is that teachers are interested in students' reasoning, to see what students know and what they do not know so that they can use this knowledge in future lectures. Also, teachers would like to see that students could perform certain operations that are more difficult to measure using multiple-choice items. Furthermore, students are sometimes in favor of open-ended questions because they have the feeling that these questions better reflect what they know.

On the statistics exam, subscores of the open-ended questions and subscores of the multiple-choice questions were reported to students during the inspection of the exam results. Although the teachers did not provide further diagnostic information from these subscores, it is not unreasonable to take the next step and to consider whether these subscores provide added value such that students may use information from the subscores to determine their study strategy for a possible re-sit exam. We used the function prmse.subscores.scales from the *R* package sirt (Robitzsch, 2016) to calculate these PRMSE's.

3.3 Results

For both exams we calculated the $PRMSE_s$ and the $PRMSE_x$. The $PRMSE_s$ equals the reliability of the observed subscore. As discussed above, the subscores provided added value over the total score if and only if $PRMSE_s$ is larger than $PRMSE_x$. For both tests, the subscores did not provide added value over the total score. Below we discuss the results for each exam in more detail.

For the test theory exam, with a total test reliability of .71, the observed subtest score reliabilities ($PRMSE_s$) equaled .58 for the conceptual understanding subtest and .52 for the factual knowledge subtest. Note that these reliabilities are low, but given the number of items and the type of questions that are being asked, they are not uncommon. Sinharay (2010) for example, reported an average operational subtest reliability of .38 for subtests with an average of 19 items. The PRMSE in estimating the true subtest score from the observed total score ($PRMSE_x$) was .80 for both the conceptual subtest and the factual knowledge subtest. Since the $PRMSE_x$ values are larger than the $PRMSE_s$ values, we conclude that reporting subscores would not be useful in this case. In this example the correlation between the conceptual understanding and factual knowledge subtest was .54, and the subscores with total score correlations were .90 and .85 for conceptual understanding and factual knowledge respectively.

For the statistics exam consisting of open and multiple-choice questions the $PRMSE_s$ was .63 for the open questions and .66 for the multiple choice questions, with a total score reliability of .77. Since the $PRMSE_x$ was .81 for the open questions and .84 for the multiple choice questions, both larger than the $PRMSE_s$ of both subtests, we conclude that reporting subscores would not be useful for this exam. Furthermore, the correlation between the subtests was .63, and the subtest-total test correlation was .85 for the open questions and .94 for the multiple-choice questions. Note that these results are in agreement with the suggestion made by Sinharay (2010) that subscore-total score correlations larger than .85 often result in subscores that do not have added value to the total score. Thus, reporting separate “diagnostic” subscores for the open questions and the multiple choice questions is not useful here.

We performed a replication study where we used a sample of 318 students who took the same exam in the same year 2014-15, only these were students who followed a Dutch-language program. Thus the exams were the same except for the language. We found very similar results as for the English-language program: $PRMSE_s$ = .66 and .67 for the open questions and the multiple choice subtests respectively; reliability = .78; $PRMSE_x$ = .79, .83; correlation between subtests was .62; subtest-total test correlation was .86 and .94 for the open questions and the multiple choice questions, respectively.

Note that this does *not* imply that the open questions do not add to the measurement precision of the total score. We can illustrate this by performing an item response theory analysis (IRT, e.g., Embretson & Reise, 2000) on the data. We fitted the two-parameter logistic model (e.g., Embretson & Reise, 2000) to the 20 multiple choice items. Furthermore, we fitted the graded response model (Embretson & Reise, 2000) to the five open-ended questions. The scores of the open-ended questions were re-coded in terms of the number of correct subtasks per item. This led to scoring four of the five open questions on a five

point scale (0, 1, 2, 3, 4) and one open question on a three point scale (0, 1, 2). The IRT models described above were fit to the data by means of the program IRTPRO (version 2.1, Cai, Thissen, & du Toit, 2011) using the default options offered by the software.

A nice characteristic of IRT is that it enables us to report the measurement precision (standard error) conditional on an examinee's score. Interesting is that if we compare the standard error of the examinees' scores, the open questions reduce this standard error and thus add to the measurement precision of the statistics exam, as shown in Figure 3.1. The test scores are now expressed on a theta metric with a mean of 0 and a standard deviation equal to 1; these theta values were strongly related to the total score (r = .93). It can be seen that across all achievement levels (theta scores) the combination of multiple choice items and open ended question resulted in a lower standard error than using only multiple choice items, especially at the higher range of the scores.

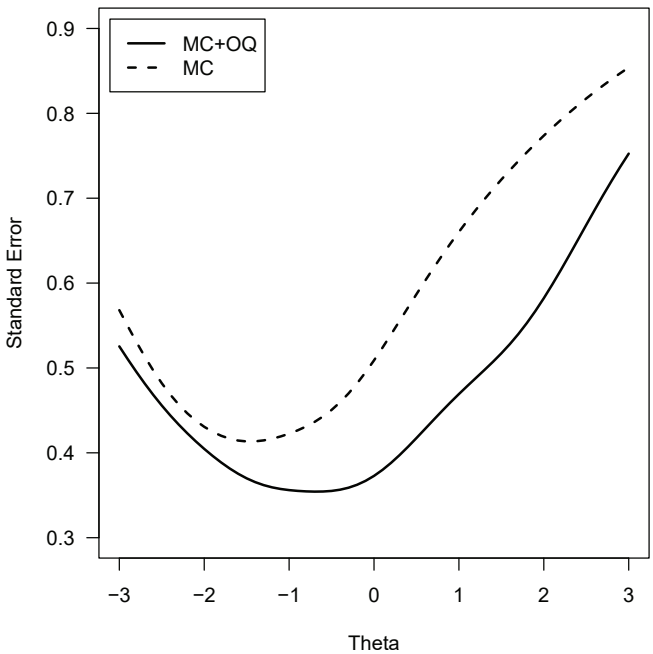


Figure 3.1. Standard error of the achievement score (denoted Theta) for the 20 multiple choice items (dashed line) versus 20 multiple choice items plus 5 open ended questions (solid line). MC = multiple choice questions, OQ = open-ended questions.

One could argue that the different number of items between both tests (i.e., 20 multiple choice items versus 25 multiple choice plus open question items) explains the difference displayed in Figure 3.1. We verified that this is not the case by a supplementary simulation analysis. The standard error of theta for all possible tests composed of any 15 multiple choice items plus the 5 open questions was computed, for a total of 15,504 such parallel tests. Figure 3.2 shows the mean standard error across the datasets, together with 95% variability bands around the mean value. It can be verified that, for theta values

above -1 the standard error based on tests including the open questions were smaller in comparison to the test based on 20 multiple choice items. Thus, the open questions do add to the measurement precision of the total score, in spite of their modest contribution to measuring the true open question subtest scores. This may partly be explained by the partial credit scoring. In general, polytomous scoring increases the test reliability as compared to dichotomous scoring (e.g., Maydeu-Olivares et al., 2009).

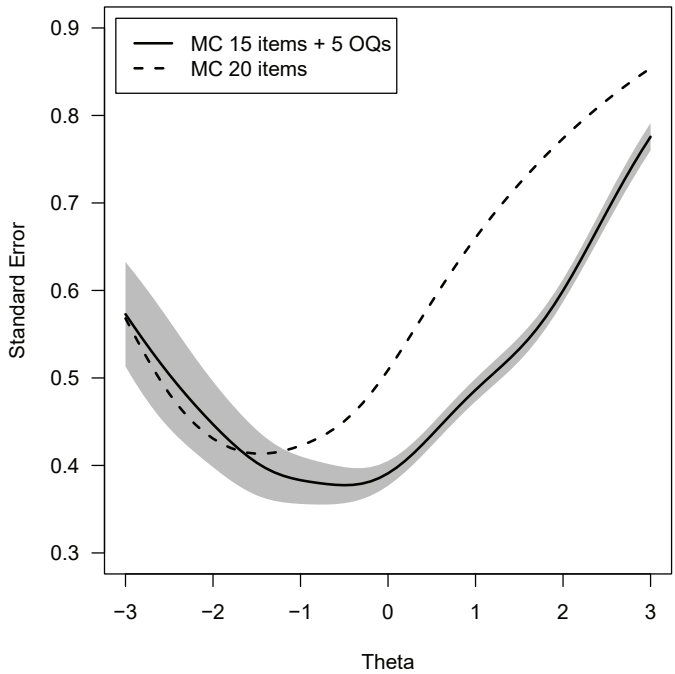


Figure 3.2. Standard error of the achievement score (denoted Theta) for the 20 multiple choice items (dashed line) versus the mean of standard error for all tests based on 15 multiple choice items plus 5 open ended questions (solid line), with a 95% variability band around the mean. MC = multiple choice questions, OQ = open-ended questions.

A further inspection of the results showed that (1) the inclusion of open questions improved the theta measurement accuracy especially for theta values between 1 and 2 (see Figure 3.2), which implies that we are better equipped to measure the score for good students, and (2) this was in particular the result of two open-ended questions that performed well to discriminate students from each other. These two items were to a large degree responsible for the improved measurement accuracy using the open-ended questions.

3.4 Discussion and Recommendations

Although most researchers and practitioners realize that unreliable subscores should not be used, or should be used with great care, in many publications in educational and psychology studies we often read sentences like “The reliability of the total score equaled .80; whereas the reliability of subscore *X* equaled .60 and subscore *Y* equaled .55.” Then, subtest scores

are being used without providing the reader any idea about how useful it is to report these subscores in addition to the total score.

We think that it is very important that teachers and, perhaps especially quality control departments that provide teachers with information about the psychometric quality of test scores, also provide information about the quality of subscores when the latter is considered important. Teachers may indicate which items form a subtest and then control departments may analyze the exam and provide feedback to the teacher. When both total scores and subscores are being reported, a teacher should show that these subscores have added value to the total score because they are interpreted as if they provide information independent from what is also reported in the total score.

Note that in our empirical examples the reliability of the total test was rather modest, which is not surprising given the small number of items in the total test. What is informative, however, is that the correlations between subscores were rather small, suggesting that they can be considered distinct (Sinharay, 2010). When subtests correlate highly, this suggests that the questions of the subtests measure similar things, and that there is a lot of shared variance. Even though the subtest correlations are not very large, however, we found in both exam examples that the estimated reliabilities based on the total tests were much larger than the observed reliabilities of the subtest scores. This was because the correlation between the subtest scores and the total scores was high: .85 and .94. This means that the subtest scores do not give reliable information about performance on that subtest.

Another important message is that when exams are not *explicitly* constructed to be able to provide scores on subtest, both in the literature and in many cases this will not be possible to use subscores in addition to total scores and report something that we did not already know using the total score. This is an important message for teachers in higher education. We are often inclined to overemphasize the information we can obtain for diagnostic purposes or for formative assessment from subscores on an exam. Thus a first take home message is that as we showed using our empirical examples, using subscores on the basis of standard exams does not necessarily add information to the total score. A second, related take home message is that it will take considerable effort to construct diagnostic exams.

Finally, it was interesting to see that adding a number of open-ended questions to the exam that were scored according to a number of well-described instructions resulted, in general, in more measurement precision than when only using dichotomously scored items. This could be explained by the scoring system: Each open-ended question consisted of a number of dichotomously scored subtasks, thus, in fact, lengthening the test with more than one dichotomously scored “item”. These results are also interesting in the light of the often-found result that open-ended questions do not add to the reliability of a test (see e.g., Hift, 2014). Perhaps when we use a well-described scoring system, these types of open questions may both increase the face validity of an exam and the reliability of an exam, at the expensive, of efficiency.

Note: Chapter 4 has been submitted as

Boevé, A. J., Albers, C. J., Bosker, R. J., Meijer, R. R., & Tendeiro, J. N. *Implementing Practice Tests in Psychology Education.*

Chapter

4

Implementing Practice Tests in Psychology Education

4.1 Introduction

Given the massification of higher education there is a continuing search for how to maintain and improve the quality of the teaching-learning process (Hornsby & Osman, 2014). With large classes of, sometimes, hundreds of students, the teacher-student ratio becomes very small, leaving teachers with little time and resources to monitor their students' learning process. Assessment is understood to be a driver of student's learning process (Gibbs, 1999; Schuwirth & Van der Vleuten, 2011). By using web-based technology, teachers can provide large groups of students with assessments for learning and feedback through, for example, practice tests.

There are still many questions, however, as to how to implement practice tests. An important issue is the extent to which the use of practice resources or tests should be completely voluntary or, whether the use of practice resources should be accompanied with performance incentives. It is also unclear at the university were this study took place to what extent implementing practice tests in a cohort of students was associated with better performance compared to a cohort that was not offered practice tests. Therefore, the present study consists of two parts: in the first study we discuss students' use of practice tests and the association between use and student performance in three different courses that differed in the amount of participation incentives. In the second study, we investigated how two cohorts of students that had access to voluntary practice quizzes performed on the final exam in comparison to a cohort that did not have access to voluntary practice quizzes.

4.1.1 Theoretical Background

An important distinction can be made between summative and formative assessment functions (Black & Wiliam, 2003; Wiliam & Black, 1996). The summative function refers to passing a judgment, whereas the formative function of assessment is meant to aid the process of learning. Examples of summative assessments are when a final exam is used to decide whether a student has sufficiently achieved the learning goals, or when a selection test is used to determine whether a candidate is sufficiently skilled to enter a particular program. In contrast, formative assessment can be considered assessment *for* learning rather than assessment *of* learning (Schuwirth & Van Der Vleuten, 2011). In practice, the boundary between formative and summative assessment is not always clear. The primary aim of practice tests for example is to improve student learning, with the focus on the formative function of assessment although there may be a summative component if the assessment counts towards a grade. Any type of official grading or participation incentive increases the stakes of a test and increases the summative function of a test. On the other hand, tests with a primarily summative function can also offer opportunities for learning when students have the opportunity to see how they performed on different parts of the assessment.

There are various theories to explain why student learning can be improved through formative assessment. The function of assessment to aid the learning process has been theorized to consist of feed-up, feedback, and feed forward (Hattie & Timperley, 2007). Feed-up can be considered the learning goal a student is working towards,

feedback is the evaluation of a student's current standing relative to the learning goals, and feed forward involves determining the next steps to be taken in the learning process. According to Hattie and Timperley (2007) the combination of addressing these questions determines the effectiveness of feedback. Furthermore, de Kleijn, Bouwmeester, Ritzen, Raemakers, and Van Rijen (2013) showed that these three aspects were all instrumental in determining the reason why students voluntarily use of formative assessment in medical education.

Research from cognitive psychology suggests that the very act of retrieving information from memory consolidates the ability to remember information, and this is known as the testing effect (Roediger & Karpicke, 2006). The implication of this finding is that any test, whether summative or formative, can potentially benefit a student's learning process if a student actively has to retrieve information from memory. The testing effect has been consistently found in lab studies, and several authors have argued that it generalizes to educational practice (e.g., Roediger & Karpicke, 2006). Studies claiming to find the testing effect in practice, however, often use evidence from practice questions identical to questions on the final exam (e.g., McDaniel, Wildman, & Anderson, 2012). This can be problematic in practice at universities where it is prohibited to administer the same questions in a practice exam and a final exam. Furthermore, Carpenter et al. (2017) showed that for three out of four exams, there was no positive correlation between the number of practice tests completed and the score on the final exam questions that were not identical to or modified versions of practice test questions.

There is empirical research that corroborates the expectation that the use of practice tests is positively associated with student performance. Carrillo-de-la-Peña et al. (2009), for example, showed that students who participated in voluntary practice test in a proctored exam environment midway through the course, performed better on the final exam compared to students who did not participate. Other studies also found that students' use of quizzes was positively associated with student performance (e.g., Angus & Watson, 2009). In attempting to improve participation of students in formative assessment, students are sometimes given incentives for participation, such as the result of formative assessment counting towards the final grade. However, Kibble (2007) showed that when student's score on the quiz counted towards their grade, almost all students completed the quizzes with a perfect score, but did not perform as well on the actual exam. In this case, the relationship between practice test use and performance on the final exam is weakened.

There may also be another explanation for why the relationship between using practice tests and student performance is not strong. Roediger, Agarwal, McDaniel, and McDermott (2011) found that children who did not use practice tests performed better on the summative tests compared to those who did use practice quizzes, suggesting that children who did not use the practice tests were effectively able to determine whether use of the practice tests would benefit their learning process. The research by De Kleijn et al. (2013), also found that an important reason why students in higher education chose not to use practice tests was that they already sufficiently mastered the course material. The advantage then of providing completely voluntary practice tests, is that it

allows students to regulate their own learning process. In contrast, however, there is also research demonstrating that students overestimate the extent to which they comprehend the material, and that they may stop studying too soon (Karpicke, 2009).

Although the implementation of practice tests is motivated by teachers aiming to improve student learning and thus performance, only a randomized controlled trial could establish such a causal relationship. In practice, and in the present study, it is both unfeasible and unethical to randomly assign part of the students to a condition with access to practice tests and assign another group to a condition without access. This means that educational research often relies on quasi-experimental research to evaluate changes to the learning environment.

Three main approaches have been taken to determine whether implementing practice tests is positively associated with students' performance. First, some research has considered the relationship between the score on practice tests and the score on final exams (e.g., Marden, Ulman, Wilson, & Velan, 2013; Kibble, 2007). When implementing completely voluntary practice tests that may be completed any number of times, however, it is unclear to what extent the final recorded score reflects the performance of a test-taking situation, or whether students filled out the answer with information on hand (i.e., without retrieving information from memory). Therefore, using the final score of the practice tests was not considered appropriate in the present study. A second approach is to examine how often or whether students use practice tests and relate this to the final score on the exams (e.g., Carrillo-de-la-Peña et al. 2009). However, it is important to consider that students who use practice tests may simply be students that are more motivated and conscientious learners than students who do not use practice tests. In meta-analyses, the amount of explained variance in college GPA that was due to motivation and study skills was found to be around 10%–15% (Robbins, Lauver, Davis, Langley, & Carlstrom, 2004; Richardson, Abraham & Bond, 2012). A third way practice test implementation has been evaluated is by considering how different cohorts of students perform, in which there is a non-randomized control group in the form of a cohort that did not receive practice tests (e.g., Dobson, 2008). What has often been neglected thus far in cohort comparison studies, however, is the equivalence of exams in subsequent cohorts. In the present study, we considered both the relationship between student use of practice tests and student performance (study 1), as well as the cohort comparison approach, to evaluate the implementation of practice tests at a university in the Netherlands (study 2).

4.2 Study 1

The goal of the first study was to evaluate the effectiveness of implementing practice tests in three courses in a first-year psychology program at a Dutch university. The courses differed in the number and types of practice tests available to the students. In one course, the practice tests were all completely voluntary, while in two other courses, there was a combination of voluntary practice tests and practice tests for which participation could count towards fulfilling the prerequisites for taking the exam, without actually being part of the course grade. The primary research question of interest was: To what extent did students use these practice tests and is the use of practice test resources related to student achievement on the final exam?

4.2.1 Method

Course characteristics

This study was conducted with students from the University of Groningen in the Netherlands, which has two bachelor psychology programs: an international program taught in English and a program taught in Dutch. The use of practice tests was evaluated in three courses that took place in the academic year 2014/2015: a course in biopsychology in the international program with 400 enrolled students, and two statistics courses, called Statistics 1a and Statistics 1b, in the Dutch program with 330 enrolled students in Statistics 1a, and 333 enrolled students in Statistics 1b. In total 265 students participated in both statistics courses. Some students took only one statistics course as a result of re-taking a course after failing it in the previous year(s).

Biopsychology. The course Biopsychology covered 15 chapters of material spread over a period of seven teaching weeks, with two lectures each week. There were two exams, a midterm exam and a final exam, each consisting of 40 multiple choice questions. The midterm exam covered the first eight chapters, and the final exam covered the final seven chapters of the course material. Students' grades were determined by the combined score on the two exams. Lecture attendance was not mandatory, and there were no required activities or assignments for this course such as practical meetings or homework assignments.

Two types of practice tests were made available as digital tests on the online-learning platform of the university known as Nestor, a local version of Blackboard (www.blackboard.com). The first type of practice test offered to students was a quiz. A total of 30 quizzes (two for each chapter) was available, each containing 15 true/false questions. The second type of practice test was a sample exam for both the midterm and the final exam. The sample exams for the midterm and final exam each contained 40 multiple choice questions. Taking both types of practice tests was voluntary and when students completed the practice tests, they received direct feedback on which questions they had answered correctly (see Figure 4.1).

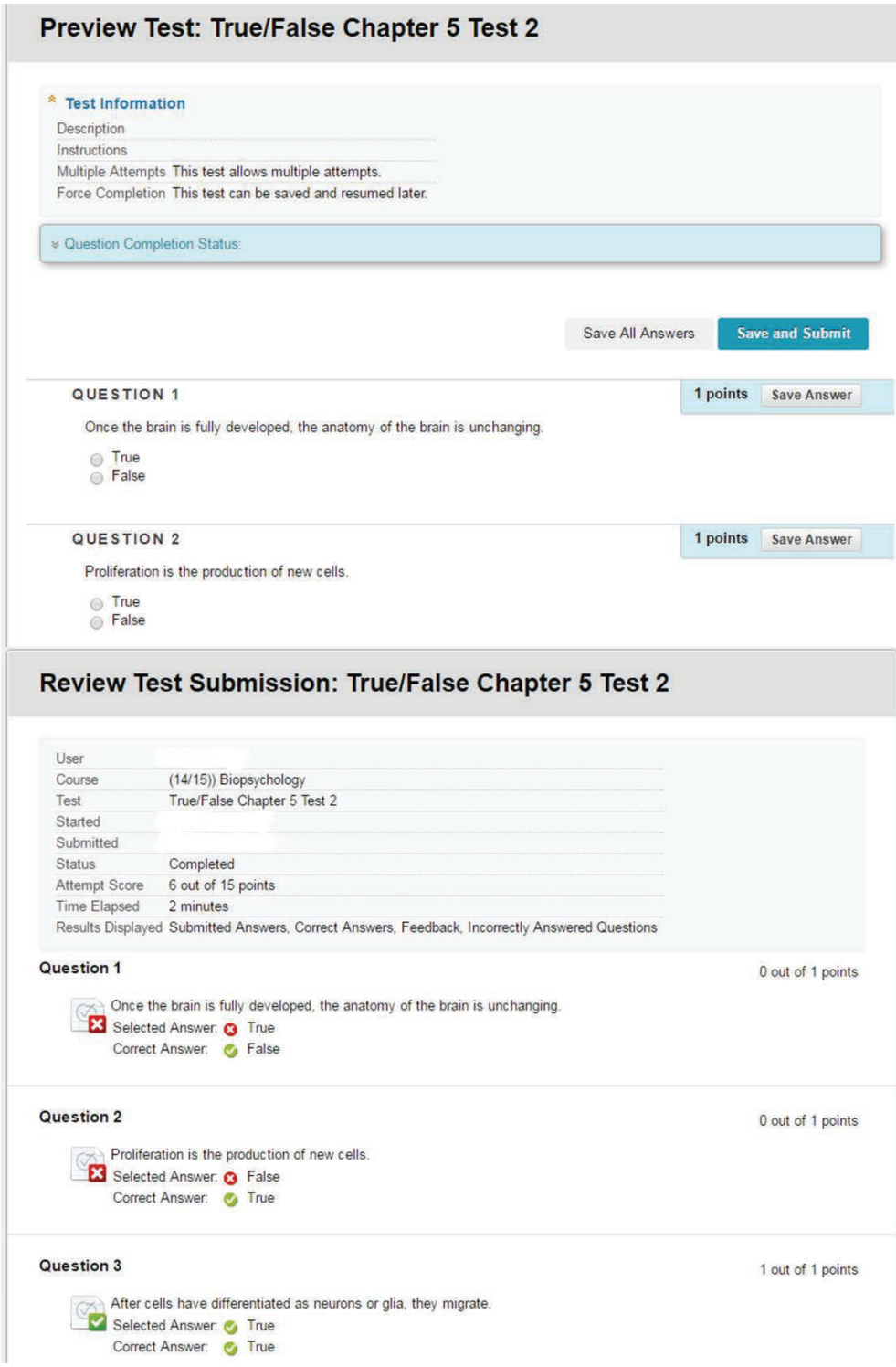


Figure 4.1. Example of practice quiz questions and feedback in Biopsychology.

Statistics 1a and 1b. The courses Statistics 1a and Statistics 1b covered five and four chapters of material from the same textbook, respectively. Both courses had seven teaching weeks, with two lectures each week of which one lecture was used to explain the material, and one lecture was used to answer students' questions. Lecture attendance was not mandatory, as in the biopsychology course. Students were required to attend a weekly practical meeting and hand in homework assignments that were not part of the final grade, but were graded as sufficient or insufficient. In addition, students taking the statistics courses for the first time were required to make use of practice test resources in a portal accompanying the textbook that offered different types of practice tests. The total amount of points possible in the portal was 2,600, and students had to earn at least 1,000 points in order to pass. There were different ways in which students could earn points in the practice portal. Students could choose to make pre-tests and post-tests of the material discussed in the chapters, they could complete tests called "learning curves", or they could answer questions about what happens when manipulating variables in an interactive environment called an "applet". This practice portal was made available and designed by the textbook authors/publisher (courses.bfwpub.com/ips8e). Its use (which practice elements of the portal were made available to students) was tailored to the learning goals of Statistics 1a and 1b by the course instructor.

A combined total of attended practical meetings, a sufficient amount of homework, and a pass for using the practice-test portal was a prerequisite for being allowed to participate in the final exam, which determined students' grades. In addition to the above course design that made use of the practice portal, there were some additional voluntary practice test resources. In Statistics 1a, there was one type of voluntary practice test, which was the sample exam covering all the material of the course and consisted of 30 multiple choice questions. In Statistics 1b, there were two types of practice tests: a sample exam, and four short quizzes for each chapter of the material. The sample exams for both courses, and the quizzes in Statistics 1b were offered through the university's online learning platform as digital tests, in the same way as the biopsychology course. These practice tests were voluntary, and students could complete the quizzes and sample exams when they wanted, and as many times as they liked. Table 4.1 provides an overview of the number and types of practice tests in each course. Students who failed the course in a previous year but did have sufficient practical attendance in that year, were exempted from the practical activities, including use of the practice portal.

Table 4.1 Amount of practice tests offered for each course, by the different types of tests.

	Chapter quizzes	**Sample exam**	**Mandatory practice portal**			
			Pre-test	Post-test	Learning curve	Applet
Biopsychology	30	2 parts	-	-	-	-
Statistics 1a	-	1	5	5	16	3
Statistics 1b	4	1	4	4	12	-

Measures

Student use of practice resources. The practice quizzes for Biopsychology were grouped in a different folder for each lecture week of the course in the online learning environment. Using the online learning environment, it was possible to evaluate the extent to which students accessed the seven folders containing quizzes for each week of the course material, as well as the extent to which students accessed the sample exams. Thus, students use of practice resources for Biopsychology was measured with two variables, one indicating the number of weeks for which quizzes were accessed (0 through 7), and one indicating the number of sample exam parts accessed (0, 1, or 2).

For Statistics 1a and Statistics 1b, there were records of students' scores on the practice quizzes, sample exams, and the number of points earned in the practice portal for each type of completed test (pretests, posttests, learning curves, and applets). Since the voluntary quizzes could be completed an unlimited amount of times and it was not possible to determine how seriously students took the quizzes, the scores on the practice quizzes were not taken into consideration. For Statistics 1a and Statistics 1b, therefore, practice test use was measured by means of whether students had completed each specific practice test (each type in the practice portal, and the voluntary quizzes and sample exams). The total number of each type of practice test completed was calculated and used for the analyses.

Student performance. For all three courses the final exam was a multiple-choice test with 4 answer options for each item. In Biopsychology, both the midterm and final exam consisted of 40 items. In order to receive a passing grade, students had to earn at least 53 out of 80 points. The exams of Statistics 1a and Statistics 1b consisted of 32 and 28 items, respectively. In order to receive a passing grade on the final exam, students in Statistics 1a had to answer 21 out of 32 exam questions correctly, and students in Statistics 1b had to answer at least 19 out of 28 exam questions correctly. For ease of interpretation, the proportion of questions answered correctly by each student was used as the measure of student performance for all three courses.

Analyses

After examining the extent to which students used the practice tests in the different courses, linear multiple regression analysis was conducted to examine the extent to which using practice test resources was related to student performance. For Statistics 1a and 1b, only students who participated in the practice portal were included in the regression analyses. All analyses were conducted in *R* (version 3.4.1; *R* Core Team, 2017). Visual inspection of the residuals gave no reason to doubt the assumption of normality (q-q plot), or homoscedasticity (plots of residuals vs. predictors). The linearity assumption may be somewhat violated for the relationship between student performance and the number of weeks of quizzes accessed by students in the Biopsychology course, leading to a potential underestimation of the amount of explained variance. However, due to the limited amount of observations at certain measurement points we restricted ourselves to the simplest model.

4.2.2 Results

For Biopsychology, out of the 440 students who enrolled, 384 attended both the midterm- and the final exam. For Statistics 1a, 305 of the 330 enrolled students attended the final exam, and for Statistics 1b, 311 students of the 333 enrolled attended the exam. There were 265 students who attended both the Statistics 1a and Statistics 1b exam, and the Pearson correlation between the total scores on both exams of these students was $r = .46$ ($p < .001$).

Practice quizzes were accessed by 90% ($N = 347$) of the students who completed Biopsychology, and 77% ($N = 297$) accessed one or both parts of the sample exam. The sample exam was accessed by 65% ($N = 199$) of students who completed Statistics 1a, and 25% ($N = 78$) of students who completed Statistics 1b. In Statistics 1b, 37% ($N = 115$) accessed the voluntary quizzes. The mandatory practice portal was used by 89% ($N = 272$) of students in Statistics 1a, and 85% ($N = 264$) of the students in Statistics 1b.

Table 4.2 shows that the use of practice resources accounted for about 24% of the variance in student performance for Biopsychology, about 5% of the variance in student performance for Statistics 1a, and 8% of the variance in student performance for Statistics 1b. For the Biopsychology course, both the number of quizzes accessed and completing the sample exam were statistically significant predictors. Table 4.2 shows that for Statistics 1a and 1b, only the use of some practice tests predicted student performance. For Statistics 1a, completing the sample exam and the number of applets accessed were both statistically significant. For Statistics 1b, only the completion of practice quizzes was a significant predictor. Table 4.3 illustrates the difference in mean exam score between students who did not access any, or accessed all of the practice tests that were statistically significant predictors of student performance in the models.

Figures 4.3 through 4.5 provide the distribution of the exam scores as a function of (a) the number of quizzes accessed and sample exams taken for Biopsychology (Figure 4.3); (b) the sample exams taken and number of applets accessed for Statistics 1a (Figure 4.4) and (c) the number of quizzes completed (Figure 4.5). These figures visually confirm the results of Table 4.2, that is, it appears that the more students access quizzes, applets, or spend time in practicing exam questions, the better their results on the exams, although the effect is small. This is especially the case for both statistics courses, where the combined effect of all these variables leads to 3% (Statistics 1a) and 6% (Statistics 1b) explained variance.

Table 4.2 Multiple regression coefficients and model results for student performance predicted by the use of practice tests in each course, with (m) indicating mandatory portal.

	Biopsychology		**Statistics 1a**		**Statistics 1b**	
	B(SE)	*p*	*B(SE)*	*p*	*B(SE)*	*p*
Intercept	66.2 (1.3)		58.2 (2.5)		55.4 (2.5)	
Quizzes	1.8 (0.3)	<.001			2.3 (0.6)	<.001
Sample exam	2.8 (0.9)	.002	3.9 (1.5)	.01	1.7 (2.2)	.43
Pretest (m)			0.1 (0.4)	.81	0.4 (0.7)	.58
Posttest (m)			-0.2 (0.5)	.71	-0.8 (1.0)	.42
Learning curve (m)			-0.1 (0.2)	.66	0.2 (0.2)	.45
Applet (m)			1.7 (0.6)	.02		
R^2/ R^2_{adj}	.24/.23		.05/.03		.08/.06	
***F*(*df*)**	58.90 (2, 381)		2.79 (5, 266)		4.42 (5, 256)	
p	<.001		.018		.001	

Table 4.3 Percentage of students who completed none and all of the different types of practice tests and their mean exam score for the statistically significant predictors of student performance in each course.

		None completed		**All completed**	
	Test-type	*N*(%)	*M*(*SD*)	*N*(%)	*M*(*SD*)
Biopsychology	Quizzes	37 (10)	56.65 (12.62)	179 (47)	68.01 (6.95)
	Sample exam	87 (23)	56.67 (11.52)	197 (51)	66.75 (7.91)
Statistics 1a	Sample exam	106 (35)	18.88 (3.70)	199 (65)	20.15 (3.62)
	Applet	123 (40)	19.23 (3.57)	10 (3)	21.20 (3.61)
Statistics 1b	Quizzes	188 (60)	16.11 (4.06)	47 (15)	19.29 (3.89)

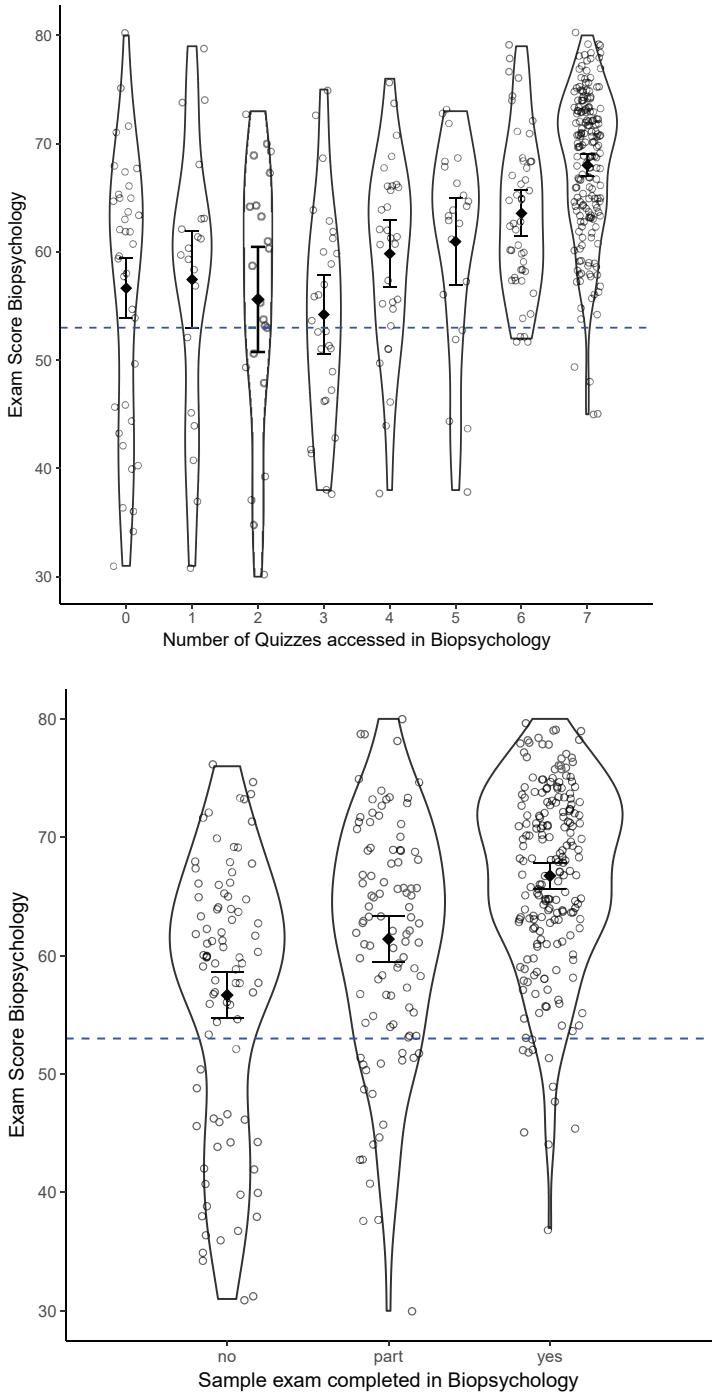
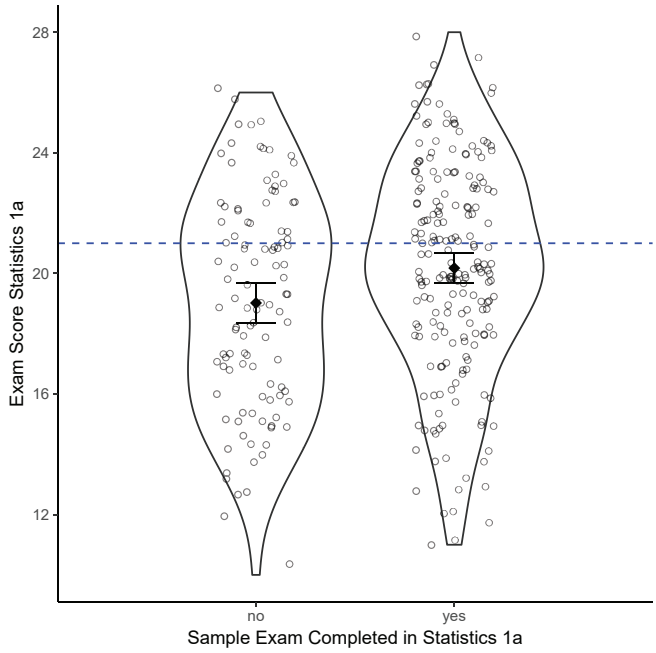


Figure 4.3. Distribution of exam scores by the number of quizzes (left) and sample exam access (right) in Biopsychology, with the dashed line indicating the cut-off score for a pass, and error bar indicating the 95% CI around the mean.



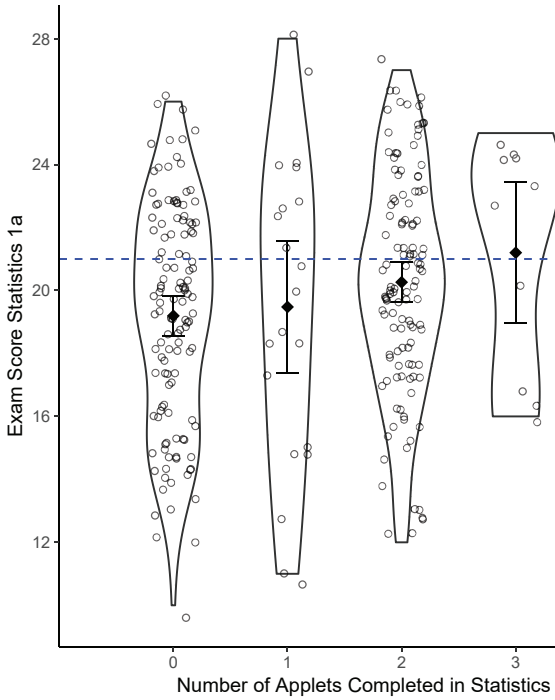


Figure 4.4. Distribution of exams scores in Statistics 1a by whether the sample exam was completed (left) and the number of portal applets (right) accessed with the dashed lines indicating the cut-off score for a pass, and error bar indicating the 95% CI around the mean.

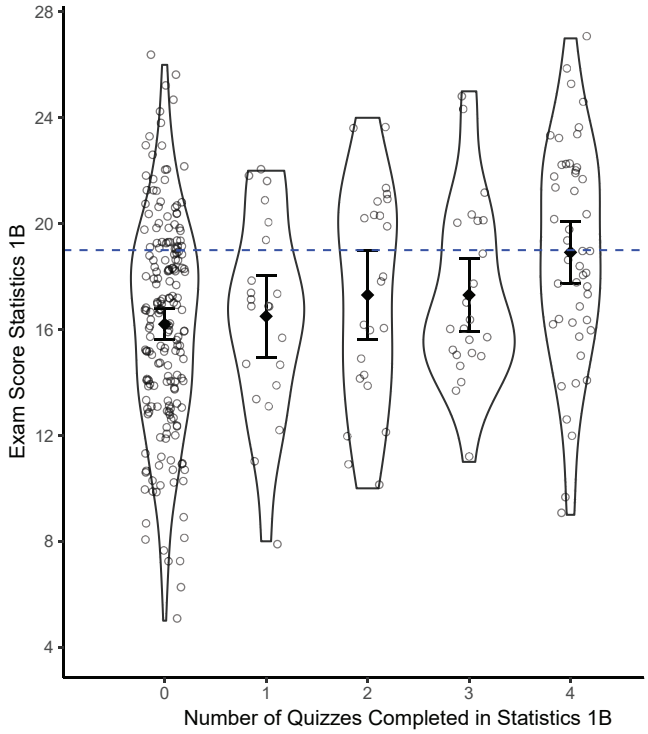


Figure 4.5. Distribution of exam scores by the number of voluntary quizzes completed in Statistics 1b, with the dashed line indicating the cut-off score for a pass, and error bar indicating the 95% CI around the mean.

4.3 Study 2

In the first study, we saw that the relationship between practice test use and student performance was strongest, yet still small, in the Biopsychology course, which also had the highest rate of participation in the practice tests. To gain further insight into whether implementing practice tests is an effective tool to improve student performance, we conducted a follow-up study for the course Biopsychology. In this second study we examined student performance across three different cohorts of which the first cohort did not have access to practice test resources, and the second and third did have access to practice tests. The primary research question of the second study was: do students from cohorts with practice test resources perform better than students from a cohort with no practice test resources? In contrast to prior research, we used item response theory (IRT, see for example, Embretson & Reise, 2000), to equate the exams from the different cohorts so that the scores between cohorts could be meaningfully compared.

4.3.1 Method

Sample. Exam results for the same mandatory course on Biopsychology as discussed in the first study were analyzed from three cohorts of students who were enrolled in the international program of the psychology bachelor program (years 2013, 2014, and 2015). Table 4.4 shows the number of students who completed the final exam for Biopsychology in each year. These numbers include students who may have re-taken the course as a result of failing the course in the past. There were 13 students who completed the exam in both 2013 and 2014, and seven students completed the exam in both 2014 and 2015.

Practice Test Implementation. The courses were designed in a similar way in all three years, except for the implementation of the practice tests. Practice tests were implemented in the international track in 2014 and in 2015 in slightly different ways. In 2014, the practice tests were made available to students via the online learning environment as images (one image for each quiz), with the answer key in a different file. In 2015, the practice tests were made available in the same learning environment in the format of digital tests that students could complete and then received feedback on (see Figure 4.1 as described in study 1). The practice tests were accompanied with an instruction for students on how to use the practice tests. As the aim of the practice tests was to improve the performance of the students, it was expected that both cohorts with practice tests performed better than the control cohort 2013.

Measures. Each exam for all three cohorts originally consisted of 80 multiple choice questions. One item from the 2013 exam was excluded from the analyses since all students answered it correctly. The final number of test items in each exam is shown in Table 4.4. Examination policy required that no two subsequent exams had the exact same questions in order to prevent cheating. Thus, to compare the total scores across the different exams in a meaningful way these scores should be placed on a common scale (that is we should equate the scores). To make this possible the exam of cohort 2015 was designed to include some items from both the exams in cohort 2014 and 2013. This inclusion of so-called anchor items enabled the scores to be equated using IRT.

Analyses. In order to account for the potential difference in difficulty between the different exams each year, the scores on the three exams were equated using the 2-parameter logistic model (2PL) with a non-equivalent internal anchor test design (NEAT; Kolen & Brennan, 2014). Based on several simulation studies, Kolen and Brennan (2014) recommended using separate calibration rather than concurrent calibration for the NEAT design, in particular when the data do not fit the IRT model perfectly. Furthermore, in order to apply the separate calibration it is typically advised to have at least a 20% overlap between tests to be equated when the total number of items is 40 or more (Kolen & Brennan, 2014, p. 288). Table 4.4 shows that this percentage is met when equating the tests of cohorts 2013 and 2014 onto the test of cohort 2015.

Typically, the newer forms of tests are equated to the oldest form of a test, which would imply equating the tests of 2015 and 2014 onto the reference test of 2013. There were, however, only 11 anchor items between the exams of 2013 and 2014, which is less than the advised 20%. Two approaches could be taken to resolve this issue: equate 2015

onto 2013, and indirectly equate 2014 onto 2013 by the path 2014-2015-2013. Another approach, however, is to equate the tests of 2013 and 2014 onto 2015. We examined the results of both approaches, and found no substantial differences. Therefore, we report the results of equating test forms 2013 and 2014 onto the reference test of 2015, since the exam of 2015 was specifically designed to include items from both previous cohorts.

Based on the above considerations, the analyses proceeded as follows in *R*: first the 2-parameter logistic model (2PL) was fit to each data set separately and item-fit was evaluated using the ltm package (Rizopoulos, 2006, version 1.0). Subsequently, the Stocking-Lord test characteristic curve transformation method was used to transform the IRT scales of the 2013 and 2014 exams to the scale of 2015 using the package equateIRT (Battauz, 2015, version 2.0-3). In particular, this provided us with all estimated abilities on the same scale, as desired. The expectation of implementing practice tests in 2014 and 2015, was to improve student performance. Therefore, we used a one-sided test for the null-hypothesis that the 2014 and 2015 cohorts did not perform better than the 2013 cohort.

4.3.2 Results

Table 4.4 shows the equating coefficients and their standard errors for each separate calibration. The density curves of the equated ability, shown in Figure 4.6, are rather similar to each other with a vertical line showing the mean ability of students after the scores were equated. We tested the one-sided null hypothesis that the difference in mean ability of the cohort of 2014 and 2015 is not greater than the mean ability of the 2013 cohort (no practice tests implemented) to answer the research question. Table 4.5 shows that this one-sided null hypothesis could not be rejected.

Table 4.4. Number of students, number of test items, number of common items with the reference cohort 2015, the estimated equating coefficients (slope *A*, intercept *B*), and mean estimated ability after equating for each cohort

	***N* students**	***N* items**	***N* common items (%)**	***A* (*SE*)**	***B* (*SE*)**	**Mean ability (*SD*)**
2015	384	80				-0.02 (0.92)
2014	348	80	41 (51%)	0.83 (0.05)	-0.07 (0.07)	-0.11 (0.74)
2013	349	79	36 (46%)	0.88 (0.06)	0.09 (0.08)	0.05 (0.81)

Table 4.5. Independent samples *t*-tests, corresponding *p*-values, effect sizes and confidence interval around the effect sizes for differences in mean cohort ability

	***t*(*df*)**	***p*-value[a]**	**Cohen's *d***	**95% CI**
INT-2015 with INT-2013	-1.12 (730.34)	.87	-0.08	[-0.23; 0.06]
INT-2014 with INT-2013	-2.70 (690.70)	> .99	-0.20	[-0.35; -0.06]

[a]*p*-value represent the one-sided hypothesis that the cohort with practice tests performs better than the control cohort 2013.

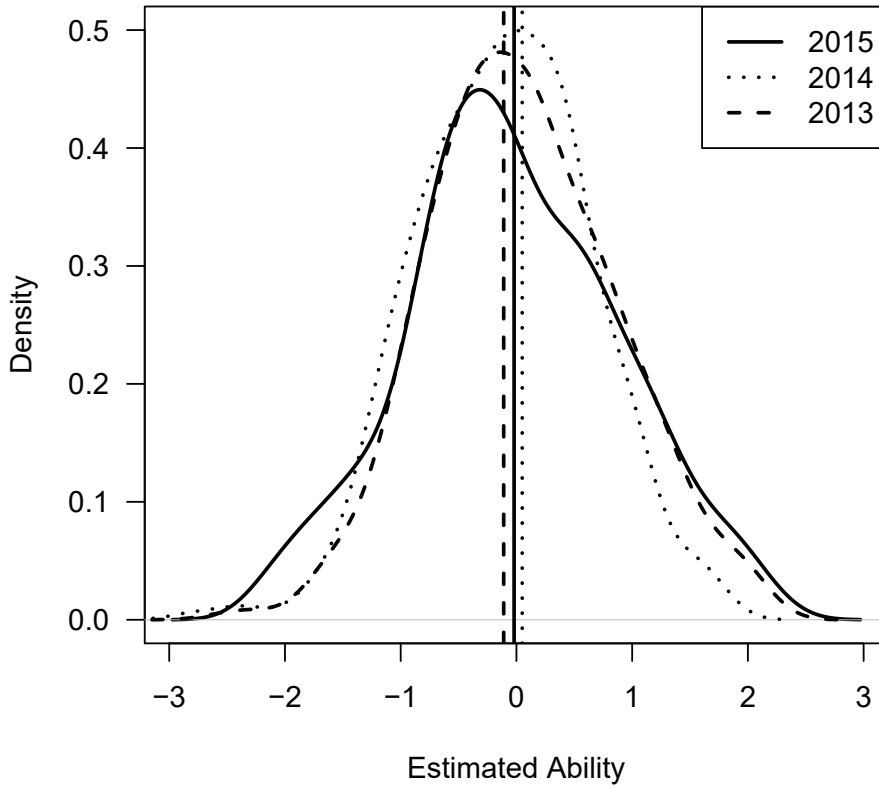


Figure 4.6. Density plot of estimated student ability for each cohort, with a vertical line at the mean ability.

4.4 Discussion

The aim of this study was twofold. First in Study 1 we sought to gain more insight into students' use of practice test resources, and the extent to which student's use of different types of practice tests was related to students' performance on the final exam in different courses with slightly different implementations of practice tests. The number of completed quizzes in Statistics 1b explained about 8% of the variance in student performance. Furthermore, we found about 5% explained variance in student performance in Statistics 1a, and a substantial amount of explained variance in Biopsychology (23%). While in all courses there was a positive relationship between the use of practice test resources and student performance, the strength of the relationship was greatest in the Biopsychology course with completely voluntary practice tests. In contrast to previous research, the incentives for using the practice tests in the statistics courses did not entail points that counted towards the final grade, but were part of a prerequisite in order to be able to complete the exam. These results seem to be in line with research demonstrating that providing incentives for use may encourage a use of tests that does not support the learning process, as suggested by Kibble (2007). Alternatively, not providing any incentives may reduce participation.

The lecturer of the statistics courses had mentioned the practice tests in the portal that accompanied the textbook, but found that these resources were not voluntarily accessed at all in previous cohorts. Thus, selecting the practice tests for students, and adding an incentive successfully encouraged their use to some extent, but not to a sufficient degree that its use was related to performance on the final exam.

In the second study we examined the extent to which cohorts in which practice tests were implemented performed better than a cohort where practice tests were not implemented. After equating the test scores, we found that the cohorts with access to practice tests did not perform better compared to the cohort with no practice tests. Thus, although there may be a positive relationship between the use of practice test and student performance (as illustrated in study 1), this does not directly translate to better student performance at the cohort level (as illustrated in Study 2).

A limitation of the present study, as in any research in educational practice, is that a teacher's goal to improve the learning process by offering practice tests cannot be causally verified to lead to better student performance. While this may theoretically be the case and a causal relationship may have been demonstrated in experimental research, it remains unclear which mechanism(s) may underlie the differential relationship between implementing practice tests and student performance. For example, student motivation and study skills have been found to explain 10-15% in student GPA (Robbins et al. 2004, Richardson et al., 2012), which is less than what was found in the present study for the Biopsychology course, but more than was found for the two statistics courses. Further research could investigate how practice test use could influence motivation and how this in turn affects subsequent use of practice tests, suggested by research on student perceptions of frequent assessment (Vaessen et al., 2016).

4.5 Conclusion

Offering practice test resources in higher education using web-based technologies is a way in which teachers can provide the means for students to receive immediate and individual feedback despite a small teacher-student ratio. Prior research, however, has used grading incentives and thus increased summative functions of practice resources (Kibble, 2007; Angus & Watson, 2009; Carrillo-de-la-Peña et al. 2009). In the present study, students' performance on the practice quizzes did not count towards the final grade, and students were given different types of resources to choose from at their own discretion.

Finally, we should realize that the relationship between the use of practice tests and study results may be a complex one. Good students may not use practice tests because they need less support in their learning process as was also found by research with primary school children (Roediger et al., 2011). On the other hand, students who do not use practice tests may lack the study skills to regulate their learning process and thus are unable to use the feedback from practice tests in a way that leads to better performance. Therefore, more insight is needed into how students' study behaviour is related to the relationship between practice test use and student performance.

Note: Chapter 5 was published as

Boevé, A. J., Meijer, R. R., Bosker, R. J., Vugteveen, J., Hoekstra, R., & Albers, C. J. (2017). Implementing the flipped classroom: an exploration of study behaviour and student performance. *Higher Education, 74*(6), 1015-1032. doi:10.1007/s10734-016-0104-y.

Chapter 5

Implementing the Flipped Classroom: An Exploration of Study Behaviour and Student Performance

5.1 Introduction

There is a continuing search on how to improve the quality of higher education so that students are able to achieve the intended learning goals. Research has shown that active learning is a promising means to this end (Freeman et al. 2014; Prince 2004). An increasingly popular method, which aims to actively engage students, is known as *flipping the classroom*. The organisation of flipped classroom courses requires students to prepare for in-class meetings, often facilitated through supportive online video lectures, and demands involvement during lectures by means of problem solving and peer instruction (Abeysekera & Dawson, 2015). A key feature that enables the change in time and place of various learning activities in the flipped classroom is the systematic use of technology both during pre-class and in class activities (Strayer 2012).

Previous research showed mixed effects of flipped classroom implementation on student performance (Davies, Dean, & Ball 2013; Mason, Shuman, & Cook 2013; McLaughlin et al. 2013; Pierce and Fox 2012; Street, Gilliland, McNeill, & Royal 2015). Tomes, Wasylkiw, and Mockler (2011) argued that in order to better support learning, educators need more insight into student perceptions of effective study strategies and how students go about studying as they prepare to demonstrate their understanding of the material in course assessment. This is especially important in the context of implementing the flipped classroom since students need to be willing to change their study behaviour. As it is unclear to what extent students comply with the changes expected from them, the present study aims to fill this gap in the literature by exploring the study behaviour of students throughout both a flipped and a non-flipped (henceforth referred to as regular) college statistics course.

5.1.1 The Benefits of Active Learning

Lectures where students passively receive information seem to be less effective than lectures where students actively participate (Prince, 2004). In a meta-analysis spanning the Science Technology Engineering and Mathematics (STEM) disciplines, Freeman et al. (2014) found an effect size of 0.47 in favour of active learning over lecture-based courses. The effect was smaller for large classes (>110 students), compared to small (<50) and medium (50–110) classes, and the effect was also somewhat smaller for studies in psychology compared to other disciplines. Furthermore, the effect sizes did not differ substantially depending on the methodological rigour of studies included, from quasi-experimental to randomized controlled trials. In active learning research, however, the implementation focus is clearly on change in the lecture setting without explicit consideration of what happens outside of the lecture setting. Therefore, it is unclear if the benefits of active learning found by Freeman et al. (2014) can be expected when the flipped classroom is implemented.

A recent study in a small class found no substantial difference in student performance between a flipped active learning course, and a regular active learning course (Jensen, Kummer, & Godoy 2015). Other research on the benefits of implementing the flipped classroom has shown varying results with different methodologies. In studies where cohorts of subsequent years were compared with medium to large groups of students

(as defined by Freeman et al. 2014), flipped classroom cohorts were found to outperform regular cohorts (Pierce and Fox 2012; Street et al. 2015). In studies with smaller numbers of students (10–50), differences in student performance were not statistically significant (Davies et al. 2013; Mason et al. 2013; McLaughlin et al. 2013), which could be due to a power problem. The advantage of cohort comparison studies is that the material and instructor may remain constant. A disadvantage, however, is that exams may be different from year to year, and in that case, difference in performance between cohorts may be the result of differences in exams rather than the implementation of the flipped classroom. In the present study it was not feasible to conduct a cohort comparison study since university policy prohibits identical exams in subsequent cohorts.

Using a different design with two groups of students following the same course in a flipped and regular format simultaneously, Tune, Sturek, and Basile (2013) found that students in the flipped course outperformed the regular course. The class sizes in this study, however, were very small for each course and insufficient information was reported for effect sizes to be computed. Although there is increasing research on student performance in the flipped classroom, there is very little research on how students study. Tune et al. (2013) found that more than half of the students indicated watching 75–100% of the video lectures. This does not, however, give insight into whether students used the video lectures as intended to prepare for class or whether it was an additional resource used to study prior to the final exam. Rather than compare the performance of students in a regular and flipped course, therefore, the present study explored study behaviour and the extent to which it is related to student performance in these two different contexts.

5.1.2 Research on student engagement

A common definition of student engagement used in the National Survey of Student Engagement (NSSE) is the time and energy students invest in educationally purposeful activities (Kuh, Cruce, Shoup, Kinzie and Gonyea 2008). However, defining student engagement and how to study it is a continuing debate (Ashwin and McVitty 2015; Kahu 2013). In the NSSE, time spent studying is operationalized as the average number of hours each week studying, with a choice of three categories (<5, 5–21, >21 hours per week). Educationally purposeful activities are operationalized as a list of activities with a Likert-type response scale ranging from *very often* to *never* (Kuh et al. 2008). The advantage of this approach to measuring how much time and energy students spend studying is that it can be compared across institutions. A limitation, however, is that it does not give insight into the varied nature of how students study throughout specific courses. Therefore, the operationalization of study behaviour as student engagement in the NSSE is of limited use in the study of change in a specific educational context like the implementation of a flipped classroom.

Another instrument measuring student engagement is the Course Student Engagement Questionnaire (Handelsman, Briggs, Sullivan, & Towler 2005). While this instrument is helpful in understanding the multi-faceted nature of student engagement, it approaches study behaviour as an individual trait that does not depend on or change throughout different courses. The same applies to instruments commonly used from

other theoretical approaches such as Vermunt's Inventory of Learning (Vermunt & Vermetten 2004), the Study Process Questionnaire (Biggs, Kember & Leung 2001; Fox, McManus and Winder 2001), and the Motivated Strategies for Learning Questionnaire (Pintrich, Smith, Garcia & Mckeachie 1993; Credé & Philips 2011). There are no validated instruments to systematically study the actual behaviour of students throughout a course in a specific context. When behaviour is the focus of change in innovations such as the flipped classroom, however, it is very important to gain insight into the mechanism that is targeted. For this reason, the present study used a diary-type instrument that was designed for the present study, similar to the approach used by Tomes et al. (2011).

In the 1980s, students' study behaviour and curriculum characteristics were studied using behaviour diaries in the Netherlands on a large scale (e.g., Van der Drift & Vos 1987). These studies showed that students concentrated their study time in the days before the exam in lecture-based courses and did not spend much time studying throughout the course, unless specific deadlines in the course required them to do so. Research on the relationship between student performance and time spent studying has yielded mixed results (Credé, Roch & Kieszczynka 2010; Dollinger, Matyja & Huber 2008; Nonis & Hudson 2006; Schuman, Walsh, Olson & Etheridge 1985). A meta-analysis on class attendance and student performance in higher education in the United States has shown that class attendance predicts almost 20% unique variance in college grade point average (GPA) over standardized achievement scores and 13% of unique variance over high school GPA (Credé, Roch & Kieszczynka 2010). It is unclear, however, to what extent prior research on the flipped classroom was conducted with mandatory class attendance (Dove 2013; Mason et al. 2013; McLaughlin et al. 2013; Tune et al. 2013). Therefore, lecture attendance, which was not mandatory, was also taken into account in the present study.

5.1.3 Student regulation of learning

The motivation of students is considered an important prerequisite for their ability to regulate their learning process. It is beyond the scope of this chapter to extensively review theories of motivation and self-determination (Deci, Vallerand, Pelletier & Ryan 1991; Niemiec & Ryan 2009), and self-regulation (Zimmerman 1990). Typically, self-regulated learning is seen as a trait that some students possess and those who score high on self-regulation perform well in any educational context regardless of how it is designed. Other research has focused on the study skills employed by students and the potential of training study skills in order to improve student performance (Hattie, Biggs, & Purdie, 1996). From this perspective, students who master study skills will be able to regulate their learning because they possess the skills to learn in an effective manner. The demands of varying learning environments, however, are often disregarded.

A line of research in which the learning environment has been included is that of the learning orientations (Vermunt & Vermetten 2004). From this perspective, students' regulation of learning can described by different typologies that are to some degree stable, as an individual trait, but also subject to change in different educational contexts. Furthermore, Vermunt and Verloop (1999) recognized that the degree of student regulation

and varying degrees of external regulation depending on specific educational contexts can affect the learning process. Vermunt and Vermetten (2004) noted that "Especially when students enter a new type of education, there may be a temporary misfit, or friction, between the students' learning conceptions, orientations, and strategies, and the demands of the new learning environment (p.280)." This seems especially relevant in the context of implementing the flipped classroom, where what normally happens in the lecture must now be done by students, and what students normally do at home is done during the lecture.

5.1.4 Research questions

The primary goal of the present study was to explore student study behaviour throughout a flipped and a regular course. Based on the literature discussed above, two main research questions were formulated: (1) How do students study throughout a flipped and a regular course? and (2) To what extent is study behaviour in a flipped and a regular course related to student performance? These main questions were investigated using quantitative data and to complement these findings with insights from qualitative data, a third exploratory question was formulated: To what extent did students in the flipped course refer to regulating their learning in the course evaluations?

5.2 Method

This study was approved by the ethical committee of the Heymans Institute of Psychology (number ppo-013-111) at the University of Groningen. In order to gain insight into how students spend their time studying throughout a course, students were invited to respond to bi-weekly online diaries. Participation in the study was voluntary, and students were informed about this at the beginning of each online diary, with a notification that by proceeding they gave informed consent for participation. This is in accordance with the informed consent policy for online data collection at the Heymans Institute of Groningen (www.rug.nl/research/heymans-institute/organization/ecp).

5.2.1 Participants

Students study behaviour was investigated in both a flipped and regular course on introductory statistics. Students in the flipped course were enrolled in the pedagogical science major, and students in the regular course were enrolled in the psychology major. While these are different groups of students, they were the most similar groups possible in terms of size and composition in the present research context. Since a cohort-comparison design was not feasible, the present design was the fairest comparison possible in an ecologically valid setting. A total of 205 students completed the flipped course, and 295 completed the regular course.

Course Design. For students in both the flipped and regular course, this was their second introductory course on descriptive and inferential statistics. The courses were taught in the first half of the second semester, February–April 2014, and covered the same material using the same book. Both courses had a different instructor, and these instructors had taught the first introductory statistics course to the same group of

students in the previous semester. In prior years, the instructors had worked together to develop the curriculum.

Students in both courses were required to participate in 7 mandatory practical meetings (in groups of about 20 students), for which they had to complete homework. The content of the practical and homework assignments were almost identical, with document analysis revealing an average of 80% exact overlap across weeks. Sufficient practical meeting attendance and handing in the homework on time was a pre-requisite for being allowed to participate in the final exam. The score on the final exam then determined students' final grade, so the exam counted 100%. Lecture attendance was not mandatory in either course. The regular course consisted of 7 lectures, whereas the flipped classroom course consisted of 13 lectures, a difference that was also present between the courses before the flipped classroom was implemented.

As part of the flipped classroom design, students in this course had the opportunity to view a 15-minute lecture-preview video. Furthermore, each student was required to hand in at least one question about the material to be covered during the lecture, for at least 8 out of 13 lectures, before the lecture took place. During the lecture, students were presented with problems (multiple choice questions). First, students were asked to provide an answer to the question themselves and use their smartphone or laptop to answer the question. Next, students were given time to discuss the answers to the question with peers, and, again, answer the question using their smartphone or laptop. Subsequently, the lecturer discussed the answers to the question, also referring to the questions that were sent in by students prior to the lecture.

5.2.2 Materials and procedure

Study behaviour. Students were invited to fill out an online diary of their study behaviour on Mondays and Fridays throughout the course. On Friday, students were asked to report on their study activities from the previous Monday through Thursday, and on Mondays they were asked to report on their study activities from the previous Friday through Sunday. For each day, the online diary contained three questions: 'did you study for statistics last {Monday, Tuesday, Wednesday, Thursday}?' If the answer to this question was no, the diary skipped to the next day. If the answer to this question was yes, the following question was 'which of the following activities did you conduct on Monday …?' and the subsequent question asked students to indicate how much time was spent on those activities selected in the prior step.

In collaboration with the course lecturers, the following study activities were included in the diary for the regular course: reading the material, summarizing the material, working on homework, completing practice (exam) questions, receiving extra tutoring, and 'other' (which could be specified by students). For the flipped course, the topics were identical, with one extra topic namely watching the online video lectures. For all the activities, students could select time slots of 15 minutes, ranging from 15 minutes to 5 hours (20 options). The diary was designed using Qualtrics (www.qualtrics.com), see the appendix for an example of the behaviour diary used in the flipped course. As an incentive for structural participation in the online diaries, 20 gift vouchers were raffled

among students who had responded to at least 80% of the diaries both half way and at the end of the course.

Lecture attendance. During each lecture, students of both courses were invited to place a check next to their name on a list to indicate presence. It was made clear to students that checking their attendance was for the purpose of research and was in no way related to assessment in the course. Halfway through the course and at the end of the course, lecture attendance by student number was published on the course website inviting students to check its accuracy and contact the researcher with corrections. Across both courses, 14 students notified the researcher with corrections in the lecture attendance list.

Student performance. The final exam consisted of 30 multiple choice questions for the regular course and 36 multiple choice questions for the flipped course. Of these tests, 28 questions were the same for both courses. Therefore, the number correct out of these 28 overlapping questions was used as the measure of student performance in this study.

Student evaluations of the flipped course. In accordance with university policy, anonymous course evaluation forms were handed out during the final exam and collected as students left. The institutional course evaluations contained three open questions that were formulated as follows: 1) How could this course be improved? 2) What were you most satisfied with in the course? And 3) what did you learn most by following this course?

5.2.3 Analyses

In order to answer the first research question (*How did students study throughout a flipped and regular course?*) the number of days studied, total time studied, and total time spent on specific learning activities was computed separately for each week of the course. In these measures respondents who did not complete both diaries for a particular week were excluded. The patterns for the specific activities tutoring and the open category other were excluded from the analyses due to the scarcity of occurrence.

In order to answer the second research question (*To what extent is study behaviour in a flipped and a regular course related to student performance?*) the total number of days studied, the total time studied, and the total time spent on different learning activities was computed over the entire course for every respondent. No respondents were excluded from these analyses, and the totals were divided by the number of days a respondent had participated for comparable scaling. Multiple regression was used to investigate whether the amount of time spent studying, the number of days studied, and the number of lectures attended explained variance in student performance. Weighted least squares regression (cf. Draper and Smith 2014, Ch. 9) was used where the weight attached to each individual's response was proportional to the number of diaries completed. This way the respondents who completed more diaries provided more information towards the regression model.

In order to answer the third research question (*To what extent did students in the flipped course refer to the process of regulating their learning in the course evaluations?*),

analysis of the course evaluations began by reading and re-reading student responses, in search of elaborations relating to the regulation of the learning process. Examining the course evaluations in this way can be considered a deductive approach to thematic content analysis (Burnard, Gill, Stewart, Treasure and Chadwick 2008; Elo and Kyngäs 2008). Due to the focus of the research question in the present study, evaluations with affective statements (like or dislike X), or opinions that were explained without reference to the learning process were excluded from the analysis.

5.3 Results

5.3.1 Response Rates

Response rates for the bi-weekly diaries for both courses are depicted in Figure 5.1. This figure shows that the response rate initially and throughout the course was larger for the students in the flipped course. A total of 78 students (26%) in the regular and 98 students (48%) in the flipped course completed at least one out of 19 bi-weekly diaries. For respondents of the first diary in the regular course the mean age was 19.5 (*SD* = 1.3), with 16% males, and for respondents of the first diary in the flipped course the mean age was 19.4 (*SD* = 1.6) with 2% male in the flipped course. An average of 11 diaries were completed by respondents in both the flipped and regular course. However, the distribution of the number of completed diaries differed, with 24% of respondents in the regular course completing one, and 8% completing all study behaviour diaries. In contrast, for the flipped course 3% of the respondents completed one, while 17% completed all diaries. Students who completed at least one study behaviour diary had an average of one more question correct on the final exam compared to students who never completed a diary (see Table 5.1).

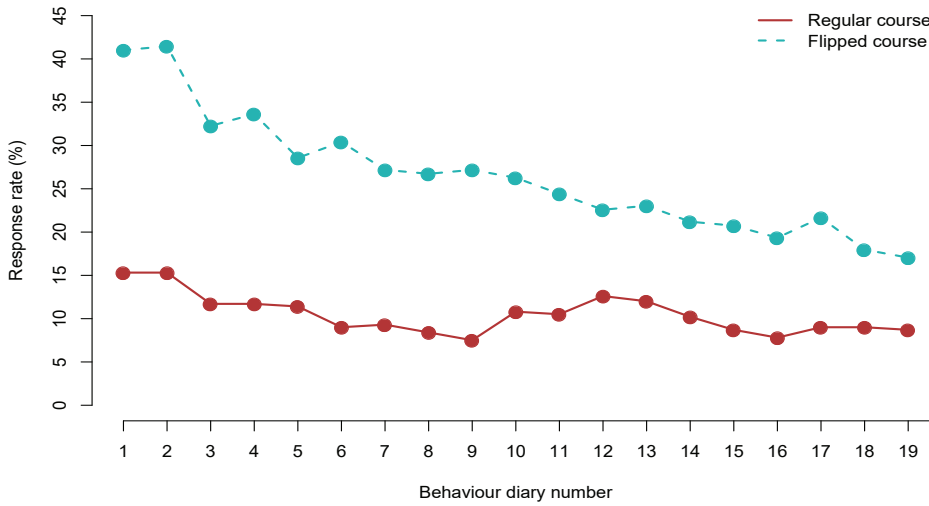


Figure 5.1. Response rates on the bi-weekly diaries for flipped and regular course.

Table 5.1. Student performance in the flipped and regular course compared between respondents and non-respondents

	N	*M*(*SD*)	95% CI difference	*t*(*df*)	*p*	Cohen's *d*
Flipped Course			[-2.4; -0.3]	-2.4 (203)	.02	-0.33
Non-respondents	107	16.5 (4.0)				
Respondents	98	17.8 (3.8)				
Regular Course			[-2.0; 0.2]	-1.7 (293)	.09	-0.22
Non-respondents	217	17.8 (4.2)				
Respondents	78	18.7 (4.0)				

5.3.2 How did students study throughout a flipped and regular course?

Figure 5.2 shows how many days and how much time students spent studying each week throughout the course. Students spread their studying for statistics over 1 to 3 days each week, while in the last week before the exam, students spread their studying over 4–5 days. Furthermore, Figure 5.2 shows that students spent no more than about 2 to 4 hours studying per week throughout the course, and in the last week 12–16 hours on average. In the first two weeks students in the flipped course spent more time studying, whereas in weeks 8 and 9 students in the flipped class room spent less time studying compared to the regular course. Overall, students' study behaviour in terms of the days spent studying and hours spent studying was rather similar in both courses. Students in the regular course who responded to the study behaviour diaries attended about 4 out of 7 (57%) lectures on average, and students in the flipped course who responded to the study behaviour diaries attended 8 out of 13 (61%) lectures on average.

Students did not spend more than 2 hours each week on average reading the course material, but spent 4–5 hours on average reading the course material in the week and a half before the exam (see Figure 5.3). Students in both the regular and flipped classroom spent less than 1 hour per week summarizing the material and studying the lecture slides. In the week and a half before the exam, students spent about 4–6 hours studying the lecture slides, and about 7 hours studying or making a summary. Throughout the course, students spent less than 1 hour per week practicing the material but in the week and a half leading up to the exam, students spent about 12 hours on average practicing the material. For the amount of time spent on homework, Figure 5.3 shows a dip in week 4 which can be explained by the fact that there was no required homework that week. Figure 5.3 shows that respondents in the regular course spent more time reading and practising in about week 8 of the course. Overall there do not appear to be many clear differences between the flipped and regular course in how students studied.

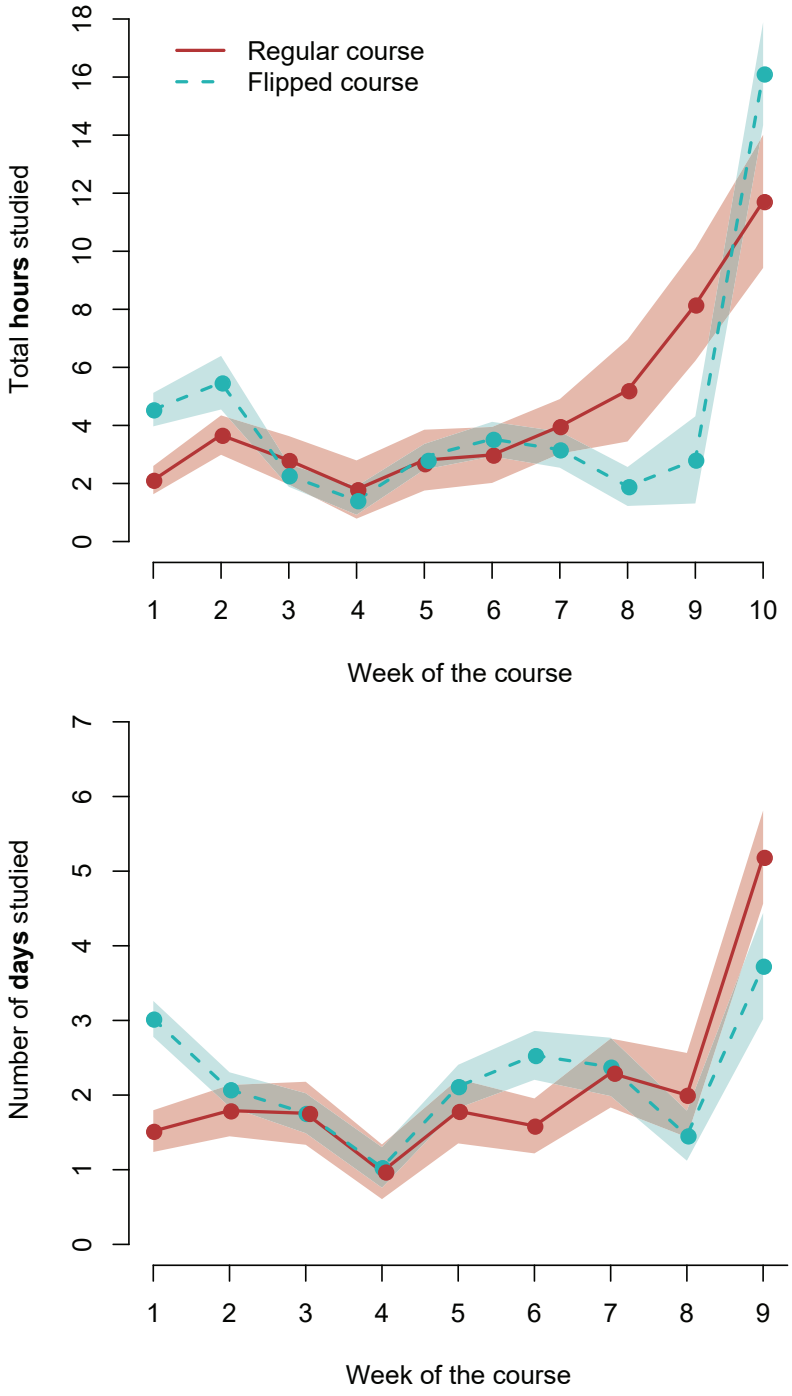


Figure 5.2. Total amount of time and number of days studied each week throughout the course with 95% confidence intervals. *Note: week 9 and 10 were combined for the number of days studied.*

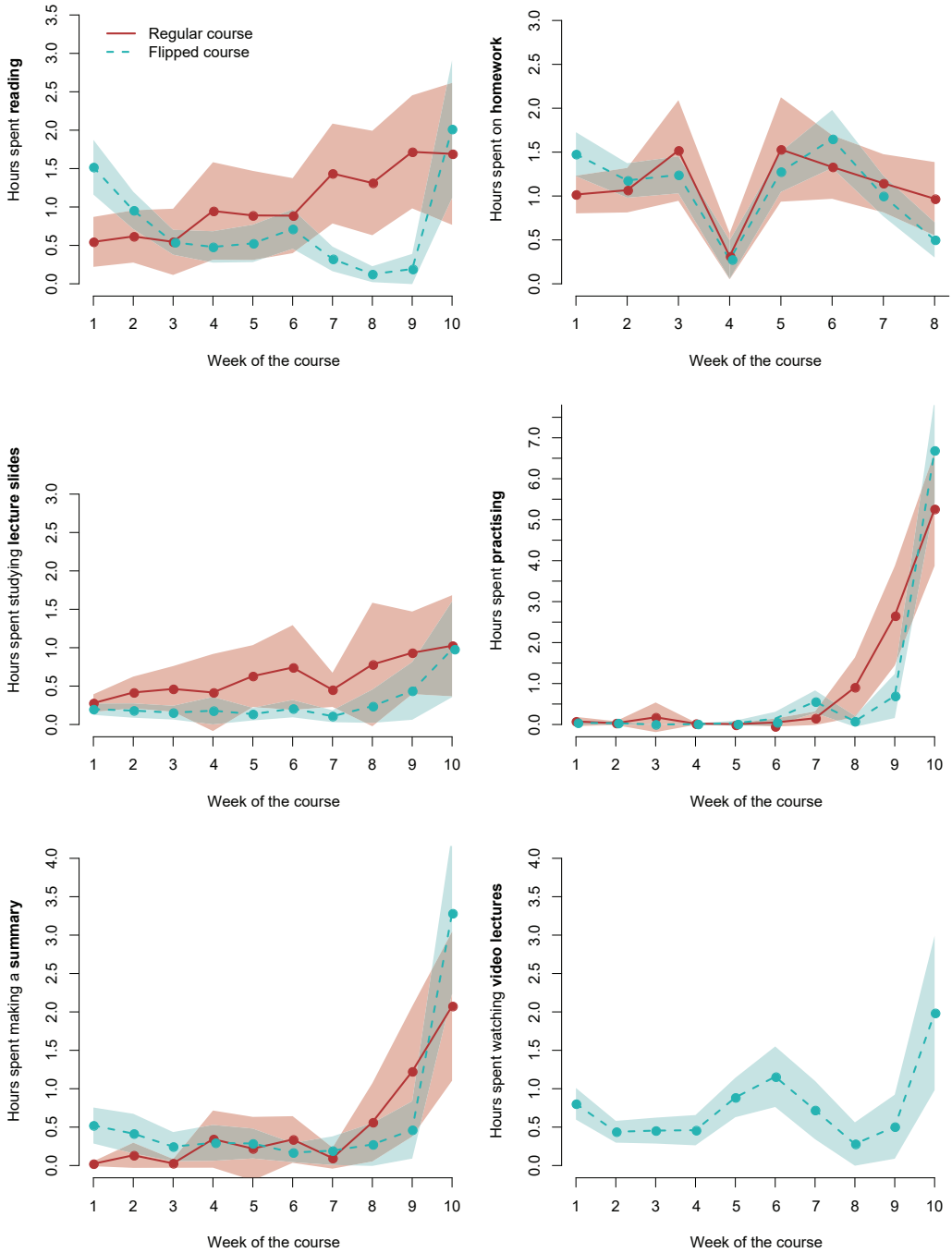


Figure 5.3. Amount of time spent on different study activities throughout the course with 95% confidence intervals.

5.3.3 How was study behaviour in a flipped and regular course related to student performance?

Table 5.2 shows the correlations between student performance, the spread of study behaviour in number of days, the total amount of time spent studying and lecture attendance. In both courses, the correlation between student performance and lecture attendance was strongest, but still fairly small (r = .23). A multiple regression model with the predictors number of days, total time, and number of lectures attended did not explain a substantial amount of variance in student performance for either flipped or regular course (flipped course: R^2 = .07, R^2_{adj} = .04, $F(3, 94) = 2.50$, p = .07, regular course: R^2 = .02, R^2_{adj} <.01, $F(3, 73) = 0.33$, p = .81). The variance inflation factors did not show problems with multicollinearity, see Table 5.3 for more details.

Table 5.2. Correlations between student performance and study behaviour (with 95% confidence intervals computed using Fisher's Z transformation). Results that are significant at the α = .05 level are displayed in bold font.

Flipped Course	**1**	**2**	**3**
1. Student performance			
2. Days studied	.06 [-.14; .26]		
3. Time studied	.09 [-.12; .28]	**.56 [.41; .68]**	
4. Lecture attendance	**.23 [.03; .41]**	.17 [-.03; .36]	.19 [-.01; .37]
Regular Course	**1**	**2**	**3**
1. Student performance			
2. Days studied	.12 [-.10; .34]		
3. Time studied	.12 [-.11; .33]	**.73 [.61; .82]**	
4. Lecture attendance	**.23 [.01; .43]**	.18 [-.05; .38]	**.26 [.04; .46]**

Table 5.3. Multiple regression results for student performance in the flipped and regular course weighted by the number of days a diary was completed

	B(SE)	**β**	***t***	***p***	**95% CI for *B***	**VIF**
Flipped Course						
Days studied	0.99 (4.63)	0.03	0.21	.83	[-8.21; 10.18]	1.51
Time studied	0.27 (0.42)	0.08	0.64	.52	[-0.57; 1.12]	1.44
Lecture attendance	0.23 (0.10)	0.24	2.29	.02	[0.03; 0.42]	1.07
Regular Course						
Days studied	4.81 (5.23)	0.16	0.92	.36	[-5.60; 15.22]	2.25
Time studied	-0.40 (0.48)	-0.14	-0.82	.41	[-1.36; 0.56]	2.26
Lecture attendance	-0.07 (0.23)	-0.04	-0.30	.76	[-0.52; 0.38]	1.11

The relationship between student performance and specific study activities was also explored (see Table 5.4). With the exception of practising in the flipped course (r = .25), none of the other study activities showed a statistically significant (α = .05) relationship with student performance. Correlations between study activities and student performance in the regular course were small and not statistically significant. Furthermore, the correlations between study activities did not appear to indicate multicollinearity in both courses, and the largest correlation was found between practising and making a summary in the regular course (r = .41).

Table 5.4. Correlations between student performance and study activities (with 95% confidence intervals computed using Fisher's Z transformation)

Flipped Course	**1**	**2**	**3**	**4**	**5**	**6**
1. Student performance						
2. Reading	-.07 [-.26; .13]					
3. Lecture slides	-.05 [-.25; .15]	.19 [-.01; .37]				
4. Homework	-.13 [-.32; .07]	**.22 [.02; .40]**	.17 [-.03; .35]			
5. Summary	.07 [-.13; .26]	**.27 [.08; .44]**	**.27 [.08; .45]**	-.08 [-.27; .12]		
6. Practising	**.25 [.05; .42]**	.03 [-.17; .23]	**.21 [.02; .40]**	-.11 [-.30; .10]	**.27 [.08; .45]**	
7. Video lectures	-.15 [-.33; .05]	.14 [-.07; .32]	**.20 [.01; .39]**	.18 [-.02; .36]	.16 [-.04; .35]	.02 [-.18; .22]
Regular Course	**1**	**2**	**3**	**4**	**5**	
1. Student performance						
2. Reading	-.02 [-.24; .20]					
3. Lecture slides	.003 [-.23; .22]	**.25 [.03; .45]**				
4. Homework	-.15 [-.36; .08]	-.08 [-.29; .15]	-.02 [-.24; .20]			
5. Summary	.14 [-.08; .35]	**.29 [.07; .48]**	.19 [-.04; .39]	-.01 [-.23; .22]		
6. Practising	.16 [-.06; .37]	**.32 [.10; .50]**	.09 [-.14; .31]	-.10 [-.32; .12]	**.41 [.21; .58]**	

5.3.4 To what extent did students refer to regulating their learning in course evaluations?

The course evaluations were completed by 173 (84%) of the students in the flipped course. Of the students who responded, 58 evaluations contained elaborations that referred to the learning regulation process. Of these evaluations, three contained elaborations referring to different aspects of the regulation process, leading to 61 comments that were split into six themes to best reflect the content of the different comments relating to the learning process. Table 5.5 shows that themes reflecting a positive experience in the regulation of learning (*video lecture supported learning, participation in lecture supported learning*, and *procrastination prevented*) were outnumbered by the amount of students with negative experiences in the regulation of learning (*more student regulation desired, more passive explanation desired,* and *student regulation necessary to benefit from lecture*). See Table 5.5 for an example comment related to each theme, and the discussion for implications these themes have for implementing the flipped classroom.

Table 5.5. Themes that emerged from student evaluations that referred to the regulation of the learning process

Code	N	Example Quote
Video lecture supported learning	8	"If I did not understand anything, or my brain was processing information, I could pause the video lecture and think about it, re-watch it, or watch it again at a later stage"
Participation in lecture supported learning	5	"The many example questions and peer explanations worked well for better understanding of the application"
Procrastination prevented	11	"by having to hand in questions I was able to keep up with the reading and was able to follow the lectures better"
More student regulation desired	10	"this way we do not have the freedom to follow our own planning"
More passive explanation desired	17	"I would have preferred more explanation of the theory in the lecture. The lectures did not contain enough explanation which is why I understood less of the material and was not able to go into the exam feeling confident"
Self-regulation necessary to benefit from lecture	10	"I think that with good preparation the lectures would be useful, but without preparation it was useless for me"

In reading and re-reading the course evaluations in search of references to learning regulation, two other themes also emerged pertaining to students experience with the flipped environment as a whole. While these did not directly answer the research question they do contribute to understanding how students coped with the change in the learning environment as a result of implementing the flipped course. The first additional theme was that students (n = 32) referred back to the design of the previous introductory statistics course to indicate how they felt the course should be designed. For example, one student

remarked “I liked the way it was in Statistics 1 because it gave me a better understanding of the material”. Secondly, several students (n = 6) demonstrated particular beliefs about who would benefit from the flipped design. A typical remark was “I think this method works well for students who are able to learn statistics easily”.

5.4 Discussion

The aim of this study was to explore study behaviour throughout a course, the extent to which study behaviour was related to student performance, and how students evaluated their ability to regulate their learning in the flipped course. By studying the time and activities of students throughout a course, student engagement was operationalized differently in the present study compared to other research on student engagement (Kahu 2013). There was no clear evidence from study behaviour throughout the course that students in the flipped course had a different study pattern compared to the regular course. The general pattern in both courses showed that students spent some time studying throughout the course, and a strong peak in time spent studying in the last week before the exam. This was also the case for the flipped course, where it could be seen that students mostly reported watching video lectures right before the exam. In contrast, the study by Tune et al. (2013) found that students reported watching 75-100% of the video lectures. With this type of retrospective behavioural question, practitioners may incorrectly conclude that students complied with the change in study behaviour asked in a flipped classroom. Thus the approach used to study the pattern of students’ study behaviour in the present study may be promising for further research into the success of implementing flipped classrooms.

The meta-analysis of Freeman et al. (2014), and much research on the flipped classroom has compared student performance in different learning environments (Davies et al. 2013; Jensen et al. 2015; Mason et al. 2013; McLaughlin et al. 2013; Pierce & Fox 2012, Street et al. 2015; Tune et al. 2013). While student performance can be compared between the flipped and regular course for the present study using the information in Table 5.1, the fairness of this comparison may be questioned due to differences in student populations, lecturers, and course design. This is a common problem for studies conducted in the real-world instead of in a controlled lab-setting. Instead, the focus in the present study was on the relationship between study behaviour and student performance in a flipped and regular course. Tomes et al. (2011) found that active learning activities such as self-testing and practising behaviour were more related to student performance than passive strategies such as reading. Although there was also a statistically significant relationship for the flipped course between practising and student performance in the present study, other correlations between study behaviour and the final exam were very small. An important difference between the present study and that of Tomes et al. (2011), however, is that they only examined study behaviour in the 10 days right before the exam. Given the expected change in students study behaviour when implementing the flipped classroom it is especially important to investigate the study behaviour throughout a course rather than only at the end of a course.

This present study is one of few to use qualitative methods to investigate student experiences with the flipped classroom. This is in line with the call for more qualitative research on student behaviour in the flipped classroom by Abeysekera and Dawson (2015). While the institutional evaluation did not contain questions about the regulation of the learning process, or even about the flipped design of the course specifically, they gave some insight into how students thought about their ability to regulate their learning in the flipped course. Six themes emerged that showed both congruence and friction in how students were able to regulate their learning in relation to the specific demands of the flipped course (Vermunt & Vermetten 2004). The first three codes in Table 5.5, *video lecture support learning, participation in lecture supports learning*, and *procrastination prevented,* can be interpreted as evidence for congruence. With the video lecture, students were able to take in the information at their own pace with the ability to pause and rewind. This can be very beneficial to the learning process compared to a lecture where all students are subject to the same pace determined by the lecturer. The interactive nature of the lecture in the flipped course was recognized to contribute to better understanding, and several students also recognized how the whole of the flipped course helped them stay engaged throughout the course. These themes demonstrate the powerful potential of the flipped classroom to contribute to the learning process.

Themes also emerged however that showed friction between student regulation of learning and the design of the flipped course. Three themes, *more student regulation desired, more passive explanation desired,* and *student regulation necessary to benefit from the lecture*, showed how students struggled with regulating their learning in the flipped course. Students expected the freedom to regulate their learning in their own way and time outside the lecture. Thus having to hand in questions about the material and prepare for the lectures each week did not always fit with their own agenda. Furthermore, a number of students indicated that they felt that the lack of passive explanation in the lectures impaired their learning process. The number of students who mentioned wanting more passive explanation was three times that of the number of students who mentioned appreciating active participation in the lecture. Given that active-learning lectures have been found to lead to better student performance (Freeman et al. 2014), this shows a conflict between what students want and what works. This demonstrates a serious challenge to educators aiming to implement not only the flipped classroom, but also other forms of active lectures. Thirdly, a number of students recognized that it was necessary to prepare in advance of the lecture in order to benefit from the lecture. Thus, while recognizing the potential of the lecture to support their learning, their own lack of ability to regulate their learning accordingly prevented the lecture from supporting their learning process.

The mix of themes showing both congruence and friction may explain why the pattern of study behaviour in the flipped and regular course, as explored in the first research question, did not appear to differ. The two additional themes that emerged also corroborate this. Students found it hard to deal with the fact that the learning environment changed compared to their prior experience. This could have led to students refusal to comply with the desired change in study behaviour that is asked of students in a flipped course. The second additional theme showed that some students did not believe they could benefit from

the flipped course design. These beliefs may have also prevented students from trying to change their study behaviour in such a way that they could benefit from the flipped course.

Although the student evaluations did not refer to the mandatory practical meetings, they could be an alternative explanation for the similarities between the study patterns in the flipped and regular course. By the presence of these additional required practical meetings the courses may have already had sufficient active learning elements, leading to a limited added value of implementing the flipped classroom as could be expected based on Freeman et al. (2014). Though Jensen et al. (2015) did not investigate study behaviour, they did not find differences in student performance between students of a flipped active learning course, and a non-flipped active learning course.

5.4.1 Limitations

The present study and that of Tomes et al. (2011) are rare cases in which students' study behaviour was investigated using diaries. A limitation of this approach is that behavioural diaries can take on many different forms, which makes it harder to use a validated instrument. The advantage of this approach, however, is that a diary can be tailored to specific contexts and research questions. As such, it was an appropriate method to investigate how students studied throughout a course in the present study.

A second limitation of using behaviour diaries was the burden on respondents. Daily measurement as in Tomes et al. (2011) would have been an extraordinary burden on students throughout the entire course, motivating the choice in the present study used bi-weekly diaries. Nevertheless, response was rather low despite the participation incentive. While dropout is not uncommon in longitudinal research (Lugtig, 2014), it may harm the representativeness of the conclusions. More research is necessary on how to measure student study behaviour in an optimal way, while keeping the burden of participation low. It is important for this methodological research to continue to take the measurement error of self-reported behaviour into account (Lugtig, Glasner, & Boevé, 2015). Despite these limitations, using diaries is a promising way to move away from general questionnaires with Likert-scales that measure trait-like attributes, towards longitudinal studies that measure actual behaviour in the context of different learning environments.

A limitation of the use of institutional course evaluations was that student perspectives on regulating their learning in a flipped classroom environment were not directly targeted in the questions. Further qualitative research focused on students' regulation of the learning process in a flipped environment may yield more and other themes than found in the present study. Since the course evaluations were anonymous, they could not be linked to students' reported study behaviour throughout the course. Therefore, further research needs to be conducted to how student perceptions about their ability to regulate learning impact their study behaviour. More research could help determine if there was a difference in the relationship between study behaviour and student performance between students who showed evidence of congruence and those who showed evidence of friction.

5.4.2 Practical Implications

This study reports the results of a localised intervention, as proposed by Abeysekera and Dawson (2015), but several recommendations for implementing the flipped classroom in large courses may be helpful for practitioners in other institutions:

1. *Consider prior history between lecturer and students*

   In this study, students following the flipped course already had a history with the lecturer. This led to expectations about how the course would be taught and how students would need to study during the course. Therefore, when implementing a flipped course for a large group of students already familiar with the lecturer, more effort may be needed to help students adapt to change. Prior history between a lecturer and students, however, may also be conducive to adapting to change particularly in courses with few numbers of students as the distance between lecturer and student may be perceived as a lot smaller.

2. *Consider the broader academic context in which the flipped course is implemented*

   Depending on how other courses in the higher education curriculum are designed, it may be more or less evident to students what type of study behaviour is expected in a flipped classroom, and what the advantages of a flipped design are. In the present study, there were no other courses implemented as a flipped course in the curriculum, and while some students clearly saw the benefits and intentions of the design, it was not so evident for many others.

3. *Expectation communication*

   Especially when teaching large groups of students, it may be difficult to gauge student beliefs about the effectiveness of particular course design, and to pinpoint when students experience friction that is destructive to the learning process. A teacher can, however, address why the flipped course can be beneficial for their learning process, and how students can benefit most. In courses with large numbers of students where attendance is not mandatory, the class composition may differ each lecture. In such cases it may be necessary to address the benefits and potential of the flipped classroom on a regular basis, to help more students recognize if they need to change their study behaviour.

5.5 Conclusion

It is important to recognize that when the flipped classroom is implemented, the demands of the learning environment change for students. The present research is one of the first to actually consider students study behaviour throughout a course in which the flipped classroom is implemented. The results do not suggest that implementing a flipped classroom is a quick fix that leads to improved student performance. Some students may benefit from the flipped design, but it may also be a source of frustration for others. More research is necessary to understand when and why implementing the flipped classroom is successful.

Note: Chapter 6 is submitted as

Boevé, A. J., Meijer, R. R., Beldhuis, H. J. A., Bosker, R. J., Albers, C. J., *Natural Variation in Grades and its Implications for Assessing the Effectiveness of Educational Innovations in Higher Education.*

Chapter 6

Natural Variation in Grades and its Implications for Assessing the Effectiveness of Educational Innovations in Higher Education

6.1 Introduction

Due to increasing performance-based accountability systems in higher education (Alexander, 2000; Liu, 2011), universities have to keep track of student performance as one of many indicators of quality and effectiveness. To achieve this, lecturers need to demonstrate that the results of student evaluations are taken seriously, and to show how changes when necessary, improve the teaching and learning environment. As a result courses are evaluated every year and lecturers keep track of how different cohorts of students perform in subsequent years. At the same time, lecturers also need to evaluate the success of implemented changes or educational innovations, where an important criterion is often the extent to which student performance has improved. This is difficult to measure in practice, however, since variation in test scores across different years may be due to different factors, including differences in exam difficulty, all sorts of cohort differences, and the effect of educational innovations. Using a Randomized Controlled Trial (RCT) to study the causal effects of an educational innovation is usually practically unfeasibly, and alternative designs are needed (Carey & Stiles, 2015; West, et al., 2008). Thus, comparing course results across years is possible, but it is not an easy task.

To disentangle different sources of variation in this context, the aim of this study was to gain insight into the amount of variation in course grades and pass-rates between years across different courses. These variations constitute "naturally expected variability", variability that is bound to exist and is not due to specific interventions. An important advantage of understanding the extent of "naturally expected variability" of exam scores is that lecturers, management, and researchers can anticipate effect sizes necessary to evaluate the success of educational changes. This is especially important in field studies in educational practice, which are often dependent on quasi-experimental designs at best. In this study we will both conduct an analysis on variation in course grades and pass rates and we will provide an example of how this information can be used in a research setting.

6.1.1 Prior Research

There is a long history of research into grading throughout all levels of education (Brookhart et al., 2016). In the early twentieth century, a lot of research focused on the variability and reliability of grades in primary and secondary education, while research on grades in higher education has focused a lot on course evaluations (Brookhart et al., 2016). There is some research on the variation of grades in higher education, mainly focused on student Grade Point Average (GPA). Kostal, Kuncel, and Sackett (2016) found evidence of GPA inflation between the mid 1990's and 2000's, and argued that instructor leniency must be an important source of the observed grade inflation. Other research on GPA in higher education focused on reliability, with Beatty (2015) finding that student GPA in the first year of college, and over the entire college period is highly reliable and did not vary much between institutions. While the focus on student GPA in research has been necessary and fruitful, research on the variability of college grades from a course perspective is lacking.

Important research has also been conducted at the primary and secondary level of education. Hollingshead and Childs (2011) showed that there was more variation over time for small schools relative to large schools in large-scale research on the percentage of students above a cut-score in Canadian primary education. School mean grades are another common

aggregate measure that is often used to consider school performance in primary education. Wei and Haertel (2011) showed that ignoring the clustering of students in classes within schools led to biased reliability and standard errors of school mean grades. In the context of secondary education, Luyten (1994) showed that there was both systematic variation in mean grades across years for specific subjects, as well as systematic variation in mean grades between courses.

The above research has important implications for the context of understanding the variability of grades in higher education. Given the more limited time, resources, and expertise of lecturers to ensure equal exam quality every year, pass rates and mean grades may vary more in higher education compared to primary and secondary education standardized testing. On the other hand, the massification of higher education may contribute to smaller standard errors given larger classes compared to primary and secondary education. The clustering of grades is an important factor to take into account as demonstrated by Wei and Haertel (2011). While research in higher education has often considered student GPA, the clustering of grades within years within courses has not been investigated. Similar to secondary education as investigated by Luyten (1994), students in higher education also take different courses taught by different teaching staff. Thus, grades in higher education are also expected to vary between courses, as well as within courses across different years.

While there is little large-scale research on course grades in higher education, course grades are often used in small-scale field studies to investigate various changes or innovations in the learning environment, with sometimes firm conclusions. Therefore, in the present study we examined the variation in course grades and pass-rates in higher education and illustrate how this information can be used to better compare course mean grades across different years.

6.2 Method

6.2.1 Data

Fully anonymized administrative records containing assessment results from the academic years 2010/2011 through 2015/2016 from the University of Groningen, the Netherlands were analyzed for the present study. The university administration provided assessment records for all first-year courses at all nine faculties of the university at that time. This research classifies as document-research for which no ethical approval was necessary according to the guidelines of the ethical committee at the University of Groningen.

Table 6.1 shows the faculties by both the full faculty name and an abridged short description that will be used in the remaining text. Table 6.2 shows the mean (*sd*) grade and pass rate per faculty. All courses from the first year of all bachelor degree programs were included. We only used first-year courses since these are obligatory and prerequisite introductory courses for further specializations later in the bachelor degree programs. Using these courses, a good picture could be obtained from the results of complete cohorts. In addition to the full cohorts of enrolled students, second- and third-year students from other bachelor degree programs may also take first-year courses in order to complete a minor. These students were also included in the data analyzed. The data analyzed had the following structure: an anonymous student-identifier, a course-code, a faculty code, date of examination, examination attempt, and examination result in the form of a grade or pass/fail.

Table 6.1 Number of assessment observations per faculty in each year, with mean (*SD*) grade and overall pass rates.

Full faculty name	Short name	*N* assess-ments	*N* year-courses	*N* unique courses	*N* unique students	Mean grade (*SD*)[a]	Mean pass rate[b]
Arts	Arts	65,798	1094	358	9,270	6.74 (0.74)	.80
Behavioural & Social Sciences	Social	73,563	427	112	8,155	6.45 (0.66)	.77
Economics & Business	Economy	83,952	354	115	9,879	6.25 (0.66)	.74
Law	Law	36,953	147	43	5,785	6.18 (0.74)	.72
Medical Sciences	Medicine	26,385	221	74	3,945	6.65 (0.61)	.80
Philosophy	Philosophy	6,301	110	36	1,388	6.73 (0.62)	.83
Science & Engineering	Science	68,209	622	139	6,709	6.67 (0.79)	.80
Spatial Sciences	Spatial	11,676	104	30	2,023	6.44 (0.65)	.71
Theology & Religious Studies	Theology	2,256	126	33	428	7.10 (0.70)	.92
Total		375,093	3205	940	47,582	6.61 (0.74)	.78

[a]Mean grade (*SD*) is computed as the mean (*SD*) of the mean grades per course
[b]Pass rate is computed as the mean of the pass rates per course

Table 6.2 Mean grade and overall pass rate for each cohort (disregarding faculty)

Cohort	Mean grade (*SD*)	Overall pass rate
2010	6.66 (0.78)	.79
2011	6.57 (0.74)	.78
2012	6.66 (0.76)	.79
2013	6.57 (0.76)	.78
2014	6.60 (0.70)	.78
2015	6.61 (0.70)	.80

In the data cleaning process, after removing empty rows and duplicate records, we selected main course results (excluding partial assessment records kept by some faculties), first-attempt results (excluding re-sits), and excluded exemption records, resulting in a total of 375,222 assessment records. Subsequently, courses were excluded if they consisted of only one student, as these have no within-course variation (n = 129). The final data consisted of a grand total of N = 375,093 assessment records from 940 unique courses (see Table 6.1 for

further details per faculty). In the appendix, tables A6(a-c) show the distribution of assessment records across faculties and cohorts, and by number of cohorts per course in the data.

The total number of students in the data equaled N = 40,087, whereas the total number of unique faculty-student combinations was N = 47,582. These numbers imply that some students took first-year program courses in more than one faculty, for example because they were enrolled in two programs simultaneously. The total number of unique student-year combinations was N = 58,612. This means that some students took courses from first-year bachelor degree programs within the same faculty in different years. Common reasons for students taking first-year program courses in multiple years include: delayed study program due to illness, unforeseen circumstances, double-degree enrollment, and following a minor-program from another bachelor program at the same faculty as the main degree of enrollment. It is important to stress that only a student's first course enrollment and assessment result were included in the data, thus there were only unique student-course combinations: a student-course combination cannot occur more than once in the data.

6.2.2 Measures

The variation in student performance was operationalized by variation in student grades and by whether students passed or failed an exam. As most continental European countries, in the Netherlands a number grading system is employed. For most courses (specific to each year), 96.8% (N = 3101) gave grades on a scale ranging from 1 to 10 where grades of 6 and higher represent a pass. Sometimes grades are given with decimals; for the present study all grades were rounded to a single integer. A small part of the courses (specific to each year) 3.2% (N = 104) only recorded whether the student passed or failed an exam, thus providing a dichotomous result.

6.2.3 Analyses

Most research on student grades in higher education has focused on student GPA, as the main outcome of interest. In order to examine the variation in outcomes across years and between courses in the present study we focused on course grades. This means that a nested structure was assumed, which is depicted in Figure 6.1. The illustration of the different nesting structures of interest to the present research on course grades, compared to research on student GPA illustrate that the same data can be assigned to different levels and that both models are essentially incomplete. In the common perspective of student GPA, the lowest level observations are not independent as each student does not take a new set of courses, but rather some students take the same set of courses. Similarly, in the present study, courses in particular years do not all have a new set of students, but rather some course-years share a common set of students. This complexity in higher education assessment data is an important challenge for researchers, but beyond the scope of the present study to solve definitively. A work-around for this problem, feasible due to the very large sample size, is as follows.

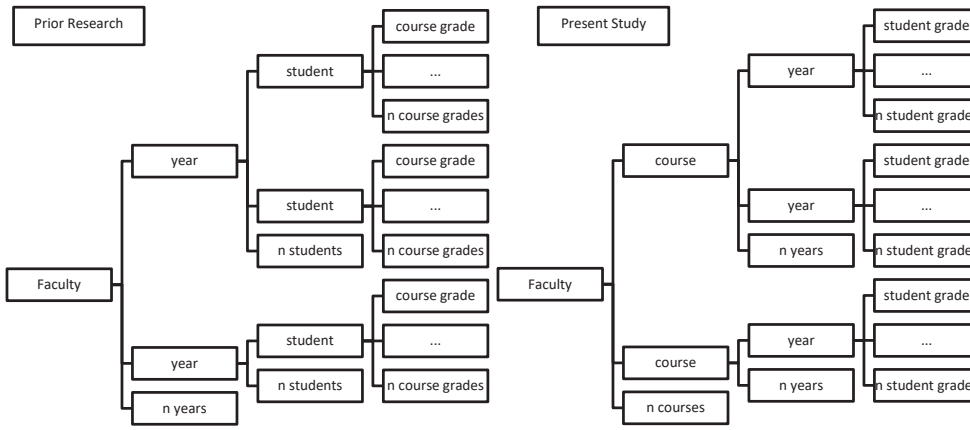


Figure 6.1. Conceptual visualization of the assumed nesting structure in prior research on student GPA (left), and the nesting structure of interest to the research question in the present study (right).

To avoid violation of independence assumptions, the analyses in the present study were repeated for 25 samples of the data where only a single assessment result was included for each student. In the first step, therefore, a single assessment result was sampled at random for each student. For students with a single assessment in a particular year and faculty, the probability of inclusion of this result would be 1. These records would therefore always be included, which may bias the findings. Therefore, a second step was added where a random 75% of the assessments selected in the first step were included.

6.2.4 Models

We constructed two models: the first model concerned the variation in mean grades and, thus, is applicable to 96.6% of the data. The second model concerned variation in the pass rate. As, obviously, a grade can always be converted into a pass/fail-statement, this model is applicable to the full data set.

Model for mean grades

The variation in course grade results was examined by estimating an intercept-only multilevel model (Snijders & Bosker, 2012; Hox, 2010) with three levels for student grades as follows:

$$Y_{ijk} = \gamma_{000} + v_{00k} + u_{0jk} + e_{ijk}, \quad (6.1)$$

where a particular grade Y_{ijk} for student i in year j in course k is modeled by the expected value γ_{000}, with a random error component for the course level (v_{00k}), a random error component for the year level (u_{0jk}) and a residual error component (e_{ijk}). All random components are assumed to be normally distributed around zero. As shown in Figure 6.1, courses are also nested in faculties. However, the number of nine faculties was too small to include as a separate level (Hox & Maas, 2005). In order to explore whether there were differences in

mean student performance per faculty, we included faculties as fixed effects, with the faculty of Arts as the reference group. In addition, we examined the proportion of variance at the year and course-level within each faculty by separately estimating the model shown in Equation 6.1 for each faculty.

The variance decomposition at different levels was investigated in the following way for student grades. First, we examined the total proportion of variance between courses and years as

$$\rho_{course.year} = \frac{\sigma^2_{u_{0jk}} + \sigma^2_{v_{00k}}}{\sigma^2_{e_{ijk}} + \sigma^2_{u_{0jk}} + \sigma^2_{v_{00k}}} \quad (6.2)$$

where $\sigma^2_{e_{ijk}}$ denotes the remaining variance in grades at the lowest level, $\sigma^2_{u_{0jk}}$ denotes the variance between years, and $\sigma^2_{v_{00k}}$ represents the variance between courses. The residuals of each level are assumed to have a normal distribution, around 0. Next we examined what proportion of the higher level variation is specific to the year level by:

$$\rho_{year} = \frac{\sigma^2_{u_{0jk}}}{\sigma^2_{u_{0jk}} + \sigma^2_{v_{00k}}} \quad (6.3)$$

Model for pass rates

To model the pass rates, a couple of additional steps were required. To examine variation in pass-rates, we modeled the log-odds of whether an assessment result was a pass (1) or a fail (0) as

$$\pi_{ijk} = \text{logistic}(\gamma_{000} + v_{00k} + u_{0jk}) \quad (6.4)$$

where π_{ijk} indicates that an assessment *i* in year *j*, in course *k* yielded either a pass or a fail which is assumed to have a binomial distribution, with an expected value of Y_{000}, a random error component across years (u_{0jk}), and with a random error component across courses (v_{00k}). After estimating this model, a second model was estimated to explore whether the mean log-odds of passing differed in each faculty. As in the analyses of grades, dummy-variables for each faculty were specified with the faculty of Arts as the reference faculty. In order to explore whether the amount of course- and year- level variance in log-odds of passing varied across faculties, the intercept only-model in Equation 6.4 was also repeated for each faculty separately.

The logg-odds are not straightforward to interpret, but can be transformed back to probabilities using the relation $p = e^{\pi}/(1 + e^{\pi})$. In each multilevel model with dichotomous outcomes, the variance of the lowest level = is scaled to 3.290 (which $\pi^2/3$, Snijders & Bosker, 2012). This means that in each model for binary outcomes using the logistic link, the residual variance is the same. To examine the variance in log-odds of passing at higher levels, the proportion can be decomposed as:

$$\rho_{year} = \frac{\sigma^2_{u_{0jk}} + \sigma^2_{v_{00k}}}{3.290 + \sigma^2_{u_{0jk}} + \sigma^2_{v_{00k}}} \quad (6.5)$$

$$\rho_{course} = \frac{\sigma^2_{u_{0jk}}}{\sigma^2_{u_{0jk}} + \sigma^2_{v_{00k}}} \quad (6.6)$$

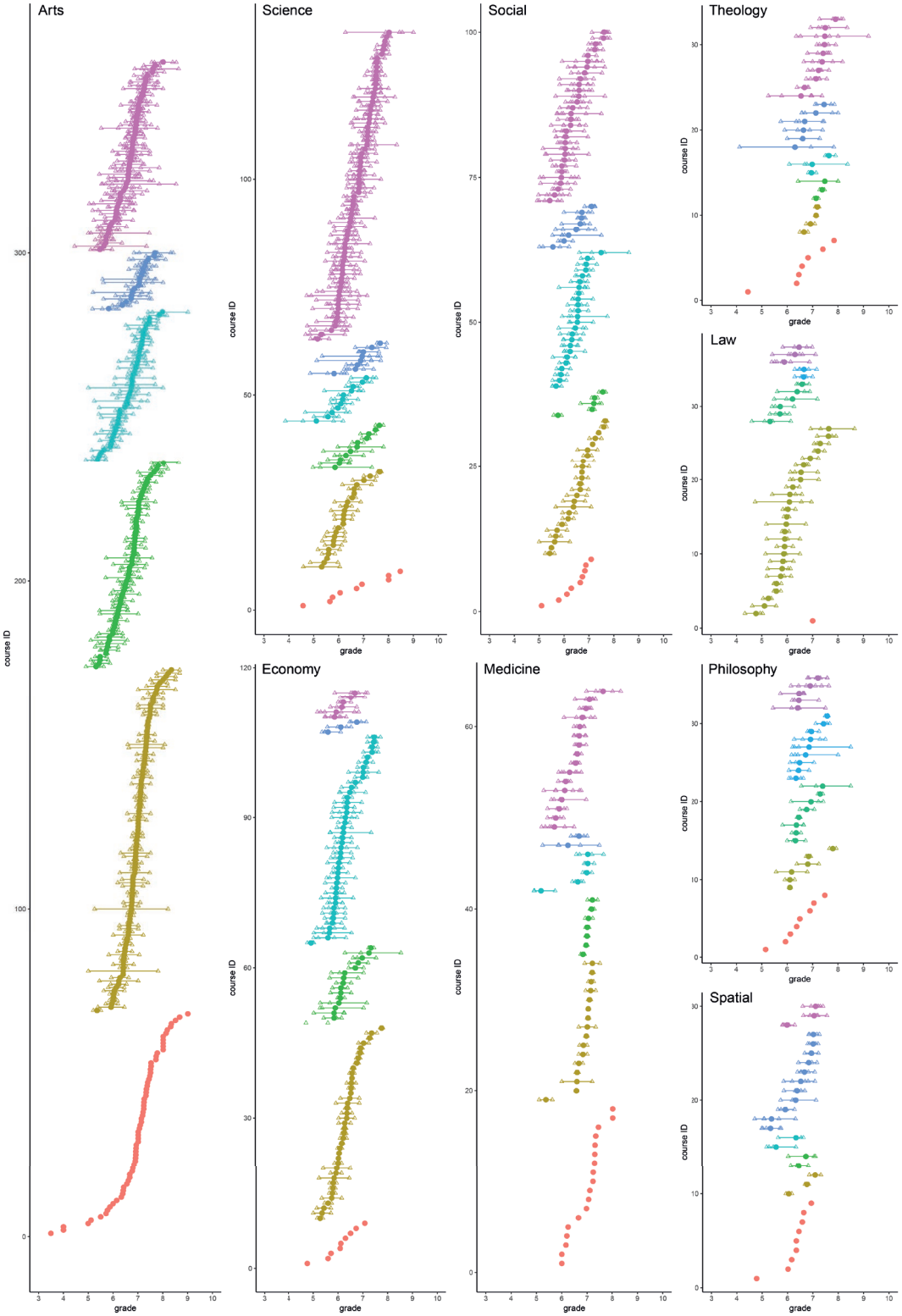
6.2.5 Software

All analyses were conducted in *R* (*R* Core team, 2017, version 3.4.1), using the lme4 package (Bates, Mächler, Bolker, & Walker, 2015, version 1.13). Full maximum likelihood estimation was used to estimate the model deviance, in order to be able to compare the intercept-only model with the model including fixed-effect dummy variables for the different faculties.

6.3 Results

To depict the variation in mean course grades Figure 6.2 shows the overall mean course grade, and the mean course grade for each year within a course for all faculties included in the data.

Figure 6.2. The variation of mean course grades within and between each course in each faculty included in the data, with colors indicating the different number of cohorts for each course (6 to 1 years from top to bottom). Each line represents the distance between the lowest mean year grade and highest mean year grade for each course, triangles representing mean grade in each year, and closed circles the mean grade for each course.


Arts
Science
Social
Theology
Law
Economy
Medicine
Philosophy
Spatial
course ID
grade

6.3.1 Course Grades

Table 6.3 shows the model results for the intercept only model, and the model with faculty included as a dummy variable in the analyses. Overall, about 17% of the variation in grades can be attributed to systematic variation between courses and years. When adding faculties as a fixed effect by means of dummy variables to the model there is a statistically significant reduction in the model deviance (Δ deviance = 107.58, $df = 8$, $p < .001$), implying better model fit. Variation in mean grades between faculties explains about 10% of the variance between courses, which is about 1% of the total variance. The size of the variance components may be underestimated due to the violation of independence, as shown in the mean variance estimates over 25 replications (see Table 6.3). Running a separate intercept only model for the different faculties shows that the amount of total course and year variation ranges between 11 to 20% (see Table 6.4). Furthermore, of the higher-level amount of variance, Table 6.4 also shows that the proportion at the year-level ranges from 25% to 52%.

Table 6.3 Estimates of the fixed effects, random effects, and model deviance for course grades

	Intercept only model	**Model including faculty fixed effect**	**Intercept only model 25 samples**
Fixed effects (*SE*)			
Intercept Y_{000}	6.59 (0.02)	6.76 (0.03)	
$D_{Theology}$		0.25 (0.11)	
D_{Law}		-0.60 (0.10)	
$D_{Medicine}$		-0.01 (0.08)	
$D_{Science}$		-0.16 (0.06)	
$D_{Economy}$		-0.51 (0.06)	
D_{Social}		-0.30 (0.07)	
$D_{Philosophy}$		-0.09 (0.11)	
$D_{Spatial}$		-0.36 (0.11)	
Random effects			Mean (*SD*)
Courses $\sigma^2_{v_{00k}}$	0.32	0.28	0.36 (0.02)
Years $\sigma^2_{u_{0jk}}$	0.15	0.14	0.16 (0.01)
Grades $\sigma^2_{e_{ijk}}$	2.27	2.27	2.48 (0.01)
Deviance	1,294,263	1,294,156	
Δ Deviance		107	
$\rho_{course:year}$	.17	.16	
ρ_{year}	.31	.34	

Table 6.4 Variance partition of grades at the different levels for each faculty

Faculty	Residual variance	Year-variance	Course-variance	$\rho_{course:year}$	ρ_{year}
Theology	1.44	0.21	0.14	.20	.41
Law	2.87	0.21	0.33	.16	.38
Medicine	1.34	0.09	0.24	.19	.29
Science	2.36	0.15	0.44	.20	.25
Arts	1.92	0.16	0.26	.18	.38
Economy	2.59	0.12	0.25	.13	.33
Social	2.23	0.15	0.26	.16	.37
Philosophy	2.31	0.14	0.13	.11	.52
Spatial	1.50	0.11	0.27	.20	.30

6.3.2 Pass rates

Based on the model with the full data, Table 6.5 indicates that about 40% of the variance in the log-odds of passing is at the year and course level. Of the higher-level variance, about 23% is due to differences between years within courses. When taking 25 subsamples of the data, so that the independence assumption is not violated the variance components are smaller. Table 6.6 shows that there is considerable variability between faculties in the amount of variance in log-odds at the year and course level, with estimates ranging from 22% to 74%. Furthermore, the relative amount of variance at the year-level within a course rather than between courses, also varies considerably, from 5% to 70%. It is important to note that these percentages of variability at the log-odds level do not translate easily to percentages at the pass-or-fail level, which will be made clear in the application.

Table 6.5 Estimates of the fixed effects, random effects, and model deviance for the log-odds of passing

	Intercept only	Model with faculty as fixed effect	Intercept only model over 25 samples
Fixed effects (*SE*)			
Intercept Y_{000}	1.80 (0.05)	1.82 (0.07)	
$D_{Theology}$		1.73 (0.29)	
D_{Law}		-0.40 (0.23)	
$D_{Medicine}$		0.21 (0.18)	
$D_{Science}$		0.11 (0.14)	
$D_{Economy}$		-0.44 (0.14)	
D_{Social}		-0.05 (0.15)	
$D_{Philosophy}$		0.04 (0.24)	
$D_{Spatial}$		-0.59 (0.27)	
Random effects			Mean (*SD*)
Courses $\sigma^2_{v_{00k}}$	1.73	1.62	0.92 (0.33)
Years $\sigma^2_{u_{0jk}}$	0.51	0.51	0.30 (0.02)
Deviance	366,947	366,885	
Δ Deviance		65	
$\rho_{course:year}$	.41	.39	
ρ_{year}	.23	.24	

Table 6.6 Coefficients for the random intercept models on the grades and log-odds of passing for each faculty

Faculty	Year-variance	Course-variance	$\rho_{course:year}$	ρ_{year}
Theology	2.60	6.94	.74	.27
Law	0.20	3.45	.53	.05
Medicine	0.14	2.24	.42	.06
Science	1.01	2.91	.54	.26
Arts	0.47	0.84	.28	.36
Economy	0.21	1.36	.32	.13
Social	0.54	1.95	.43	.22
Philosophy	0.24	0.70	.22	.25
Spatial	1.40	0.60	.38	.70

6.3.3 Application

Consider the following scenario, with intentionally simplified numbers: A course instructor is interested in implementing a new teaching method. It is not possible to do a randomized experiment, and the instructor would like to compare the results of the previous-year, that is the results prior to the implementation of the new teaching method, with that of the current year, that is, the results after implementing the changes. In both years, n = 50 students participate, and the GPA for both years is 6.00 and 6.50, respectively. In both years, the standard deviation of grades is 1.00. A standard t-test shows that the increase in GPA is highly significant $t(98) = 2.50$, one-sided $p = .007$). Concluding that, thus, the new teaching method is beneficial is misleading, as the regular year-to-year variations are not taken into account. To infer a significant increase in GPA after an educational intervention, the increase in GPA should not just be significantly above zero, but significantly above regular values obtained from year-to-year variation.

The variance partitioning of grades and year-variation in the present study can be informative: based on the estimated proportion of variance across years, a course-instructor can estimate the 95% CI around the difference between two cohort mean grades as follows:

$$0 \pm t^* \times \sqrt{(\frac{1}{n_1} + \frac{1}{n_2}) \times (\sigma^2_{year} + \sigma^2_{residual})} \quad (6.7)$$

where n_1 and n_2 are the number of students participating in the course both years, t^* is the critical t-value with $n_1 + n_2 - 2$ degrees of freedom. The course-level variance is excluded here since the result in both years is for the same course. For a random course the year-level variance component of the overall model can be used based on the intercept-only model. It is also possible to use a faculty-specific variance component if the faculty is known. Figure 6.3 shows the 95% confidence interval around the mean grade for different possible numbers of students in each cohort, based on the estimated variance components of the overall model. From this figure, it is clear than an increase of 0.5 in GPA for a course with 50 students per year is non-significant. For larger courses, for example, with 200 students per year, a 0.5 increase would be a significant effect.

Similarly, Equation 6.7 can also be used to estimate the 95% confidence interval around the log-odds of passing. In contrast to the application of mean grades this is, however, dependent on the intercept (i.e. the log-odds of the average pass grade), while the application for grades is equivalent regardless of the mean expected grade. Figure 6.4 shows the same 95% confidence interval after transforming the log-odds interval back to the probability of passing. Say a lecturer observes that the original cohort had a pass rate of .86, and observes a pass rate of .90 in the course with the new lecture method, Figure 6.4 shows that you need at least 150 students per year for this difference to be significant at the 5% level.

To illustrate how the confidence interval of the log-odds varies depending on the expected intercept, Figure 6.4 shows the interval for different possible numbers of students given three different intercepts, based on the quantiles of pass-rates in the present data. This figure shows that whether a certain increase from year 1 to year 2 in pass rate is

significant depends on the pass rate of year 1. For instance, in a course with 100 students, a 5 percentage point increase from 60% to 65% is not significant, whereas the same increase from 90% to 95% would be. In general, for pass rates closer to 1 (or to 0), smaller increase in pass rate can be significant than for pass rates closer to 50%.

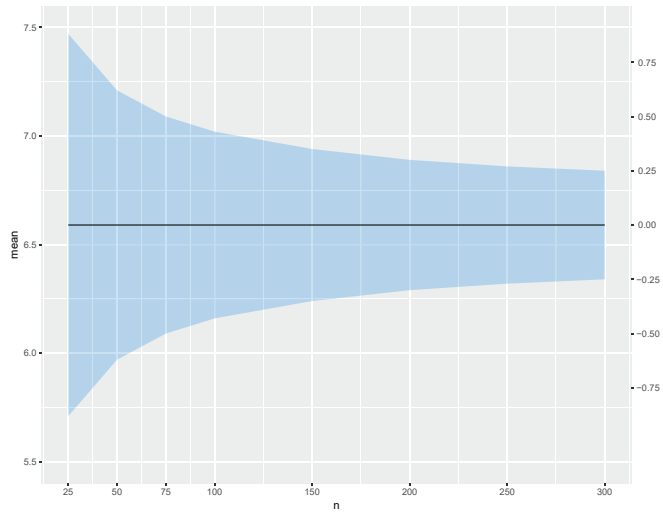


Figure 6.3. The 95% confidence interval around the predicted mean grade when $n_1 = n_2 = N$. The horizontal line is placed around the observed mean grade in the data set (left vertical axis), but the width of the confidence interval would be the same when placed around another mean value. The right vertical axis displays the deviation from this mean value.

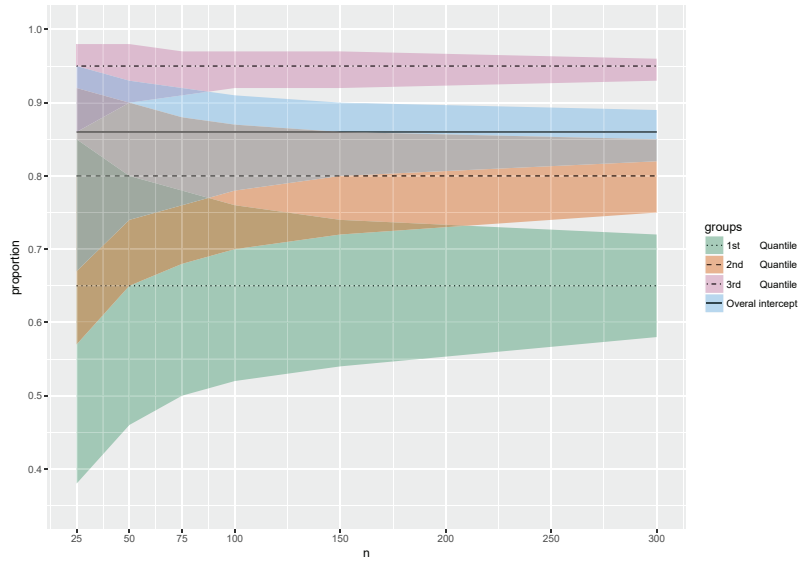


Figure 6.4. The 95% confidence interval of the probability of passing based on the overall model intercept (mean of .86), and the quantiles of mean pass-rates, when $n_1 = n_2 = N$.

6.4 Discussion

The aim of the present study was to explore the extent to which assessment results, both in terms of grades and passing, vary between years within courses and the extent to which they vary across courses within different faculties. Disregarding the different disciplines of the different courses, the present study found that about 17% of the variation in grades was at the year- and course- level. Of this variation, about 30% was due to variation within courses across different years, whilst the remaining 70% was due to systematic variation between courses. Despite the high reliability of student GPA as demonstrated by Beatty et al. (2015), the present study showed that year-over-year variations in grades may be considerable.

When examining the log-odds of whether an assessment result was a pass or a fail, we found that approximately 40% of the variance was at the year- and course- level, with 25% of this variation across different years within courses, and 75% between different courses overall. When accounting for different disciplines (faculties) in the data, the amount of variation between courses decreased slightly from 17% to 16% in terms of course grades, and from 41% to 39% in the log-odds of passing.

In line with the findings of Luyten (1994) in the context of secondary education, the present study found that the proportion of course-level variation was larger than the variation within courses across years. However, exploring discipline specific differences in the amount of variation at the course- and year-level revealed substantial differences between faculties. The overall amount of higher-level variation varied from 11%-20% concerning grades, and ranged from 22% to 72% for the log-odds of passing. Of the higher-level variance, the proportion of variance across years ranged from 25%-52% for grades, and for the log-odds of passing, the proportion of variance across years relative to the higher-level variance ranged from 5%-70%.

The implications of the findings in this study are severe, as was shown in the application section. In educational literature, innovations are often judged effective based on a direct comparison of two cohorts, without taking this the “naturally expected variation” into account. Disregarding the general fluctuation in course grades over time leads to a severe increase of false positives as innovations may incorrectly be labeled as effective. Whereas for a course with 50 students a difference in grade mean of 0.5 *SD* before and after intervention would be considered highly significant (p = .007) when disregarding this variation, the difference actually is non-significant at the α = .05 threshold. At least 75 students are needed to get the p-value below.05.

In line with the findings of Hollingshead and Childs (2011), as the number of students increase, the uncertainty around both the mean course grade and pass-rates decrease. This study demonstrated that, even with large sample sizes, conclusions about cohort differences should be taken with caution. For instance, for a large course, with 300 students per year, an increase in pass rate from 65% to 70% is not even significant. When ignoring the natural variation, this difference would be highly significant.

As evaluations of educational innovations ignore this natural variation, it is to be expected that the number of false positive findings is very large. A practical recommendation to avoid this, is as follows. Based on the number of students in a course, one can use Figure

6.3 to find the value δ which is the maximum value of the difference in mean grades in two consecutive years, $m_2 - m_1$, which would be non-significant. For instance, with $n = 50$, $\delta = 0.62$. Rather than testing for a significant difference between both means (H_0: $\mu_2 - \mu_1 = 0$, with the standard *t*-test), one can then test whether the difference between both means is significantly larger than δ or not (H_0: $|\mu_2 - \mu_1| < \delta$). For this, one can use equivalence tests (Schuirmann, 1987; Lakens, 2017). To claim a successful educational innovation, the difference between grade means should significantly exceed δ, rather than just significantly exceeding 0. When the interest lies in the pass rate, rather than the grade mean, a similar approach can be employed using Figure 6.4.

6.4.1 Limitations

The present study was focused on assessment in higher education. As always in data analysis, not all potentially relevant variables are measured. Some faculties offer multiple bachelor degree programs, and there may be systematic variation between bachelor programs within the same faculty. As type of bachelor program was not recorded in our dataset, we could not take this level into account in our analyses. Also the effect of individual lecturers could not be taken into the model as this information was not part of the data set.

Another limitation in the present study was that it is unknown to what extent courses were taught in the same way or by the same lecturers in different years. Major education innovations did happen, but not without assigning a new code to the course (thus treating both courses separately). The variance across years however, likely does include lecturer experimentation with perhaps new technology or assessment methods. Note that the average grade did not increase significantly over the years (Table 6.2).

The main limitation of this study has to do with the generalizability of the results. The present study examined the grades and pass-rates of first-year courses in higher education at a single university, the University of Groningen. Given the large amount of information, the estimated variance components could be informative for other institutions, especially those using a number grading scale. Although it is unknown to what extent the numerical findings in this study are representative for other universities, it is obvious that also at other places a considerable part of grade variation can be labeled as 'natural variation'. Thus, the message that many 'significant' findings in assessing educational interventions are actually false positives holds, but further research is needed to assess *how* many of these findings are false.

Furthermore, higher education institutes can employ the model introduced by us on their own assessment records. If they find more natural variation at their institute than we did in our study, an even larger grade increase is required for a successful intervention. Reversely, with less natural variation, smaller increases can be labeled as successful.

Finally, the present study demonstrated that assigning observations to different levels is sometimes not straightforward (see, e.g., Hox, 2010). For research on student grades in higher education, the focus has often been on the student with the interest in explaining why individuals differ in their achievement, here the focus was on how courses differed in achievement across different years.

6.4.2 Conclusion

The goal of this study was three-fold: (i) introducing a model for assessing "natural variablity" in grades in higher education, (ii) estimating the parameters in this model based on a large (n = 375,093) data set from a single university, (iii) showcasing the consequences of ignoring this natural variation in studying whether an educational intervention yields a significant increase in grades and/or pass rate. The assessment records of higher education institutes contain valuable information when examined from a course perspective rather than from a student perspective. Understanding the variation in course results across years can help lecturers and institutions to evaluate the impact of innovations at a cohort level, while reducing the risk of false positives when grades between two subsequent cohorts are compared.

Chapter

7

Discussion

7.1 Introdcution

As discussed in chapter 1 this dissertation was driven by challenges and dilemmas faced in classroom assessment in higher education. Because the research was conducted in real settings with different possibilities and limitations for research designs, the collection of studies in this thesis were based on different theoretical lenses and used different methodologies. Whenever, due to the nature of the problem, fully experimental designs (RCTs) are not feasible, employing a combination of research methodologies to answer a set of related research questions is advocated (e.g., Carey & Stiles, 2016).

7.2 Summary of the main findings

By means of a field experiment, in chapter 2 I demonstrated that students' performance on exams consisting of multiple choice questions did not differ depending on the mode of examination. The computer-based exam facilities at the time warranted an experimental design in a real-life context, allowing strong conclusions. Given the important role of technology in society, it was somewhat surprising that results of a survey among both a Dutch and an international cohort of students did not find support for computer-based exams; about half of the students still preferred a paper-based exam. Initial results of a qualitative enquiry into the reasons why students preferred a certain mode of examining showed that universities can take measures to reduce the stress students experience in such a high-stakes situation. For example, students prefer to change the layout of the exam according to their own preference (one question at a time, or all questions at once), and they like to have note-taking and other editing functionalities in computer-based exams.

In Chapter 3 I investigated whether the use of subscores on an exam can add additional information to the total scores on an exam. This is a relevant question in exam practice because there is a vivid discussion to what extent assessments can be used to guide or steer students' further learning based on identifying students' strengths and weaknesses. In assessment practice at universities, information about subscores on an exam could, for example, direct students towards parts of the content that they have not mastered so that those who failed can use this information to study for a resit exam. Based on this discussion in practice as well as recent literature (e.g., Sinharay, 2010; Sinharay, Puhan, & Haberman, 2011) the focus in chapter 3 was on the reliability of subtest scores. Although providing students with feedback about their performance based on subscores seems to be a good idea, subtest scores may not be as reliable as the total score of an exam. This was demonstrated both using an exam testing different types of knowledge and an exam consisting of both open - and closed questions. Interesting was that some of the open questions contributed to a better measurement precision, especially the open questions that used more structured scoring (and are thus most straightforward to grade) contributed least to the measurement precision.

The research presented in chapters 4 and 5 involved the use of online technology to improve the learning process. In chapter 4 the implementation of practice tests was explored in different contexts. In the first part, the relationship was investigated between students' use of practice test resources and exam results in two statistics courses and a biopsychology course. Results showed that there was a positive relationship between the

voluntary use of practice tests and performance on the final exam, but the strength of the relationship varied in the different courses. In the second part of this study the performance of cohorts with practice test resources was compared to a cohort without practice test resources by means of test equating. The mean performance of the cohorts with voluntary practice test resources was not higher than the mean performance of the cohort that did not have access to voluntary practice tests.

In Chapter 5 I explored students' study behaviour in a flipped classroom course and a regular course by means of bi-weekly diaries. The aim of implementing the flipped classroom was that students would study more, and more intensely, throughout the course, due to required preparation and active engagement during the lecture. Results from the diaries showed that students' study behaviour in the flipped classroom course was not very different from students in a regular course. Furthermore, study behaviour did not appear to be strongly related to student performance in both the flipped and regular course. Exploration of students' study behaviour in the course evaluations showed that some students experienced the flipped course design as intended to support their learning process. Other students however, demonstrated resistance to changing their study behaviour even though changing study behaviour is expected in order to benefit from the flipped classroom.

Chapter 6 was inspired by the literature on higher education and research conducted in chapters 4 and 5. Researchers in higher education often cannot use randomized controlled trials to evaluate innovations, and, therefore, often use quasi-experimental designs to compare the results when similar courses are followed, or to compare the results of the same course across different years. Student performance in this type of research is compared without taking into account the natural variation in exam scores. Based on all the grades from first year courses over a period of six years, I studied the variation in mean course grades. Overall, about 17% of the variation in grades could be attributed to the year and course level, while almost 40% of the variation in passing a course could be attributed to the year and course level. Using this information I illustrated that a statistically significant difference in course grades may fall within the expected variation in grades. Thus, observing a moderate increase in student performance after the introduction of an educational intervention may not be indicative of the intervention's effectiveness.

7.3 Limitations

The research in this thesis was conducted at a single university in the Netherlands. Therefore, an important limitation was that the results found in the various chapters may not generalize to other programs or higher education institutes, such as other types of higher education (e.g., Higher Professional Education in the Netherlands), or institutes in other countries. However, the issues addressed in the present study do reflect challenges many institutes, in different countries, face.

When implementing an innovation in education, it is often not possible to use an experimental design and, therefore, it is often difficult to evaluate the causal effect of an implementation. Perhaps to the frustration of practitioners, much research, including that presented in chapters 4 and 5, is by design unable to suggest that implementing a flipped

classroom or practice tests lead to better student performance. Perhaps field experiments need to be employed on a much larger scale in order to evaluate the impact of innovations in higher education. This could be achieved in a large well-coordinated project where randomization at the course level is possible. For example, one or more institutions may be involved, with instructors willing to implement an innovation, such as the flipped classroom. A risk in field experiments is that of spillover-effects, that is, effects of the implementation trickling into the participants assigned not to receive the implementation, so that such a design would need to ensure that only courses from a single program of study would be included. Such a large-scale design would be the most convincing, and potentially most generalizable. A sufficiently large sample size at the course level would also be imperative in order to be able to analyse the results adequately (Hox, 2010).

7.4 Scientific contributions

The aim of this dissertation was to empirically evaluate several assessment innovations in higher education. In doing so, the various chapters contributed in different ways to both methodological and theoretical debates in higher education research. With respect to the methodology, in chapter 6 it was discussed that many research projects in higher education innovations focus on the quasi-experimental comparisons of courses that are similar in content, or use subsequent cohorts of students following the same course. If the grade variation of many higher education institutes was known, then perhaps it would be possible to determine how large the beneficial effect of an educational measure should be in order to be larger than a regular variation to be expected purely based on random year-to-year and course-to-course variation. This could greatly enhance the discussions about what kind of effect size is reasonable to work towards and is needed in terms of the expected benefits of the educational innovation. It is reasonable to expect that in order to make causal claims such as “this educational intervention has the desired effect”, more stringent demands should be met compared to correlational claims such as “the mean pass grade after the innovation was statistically significantly higher than the mean pass rate before the innovation”. Other methodological contributions could be found in chapter 3 and in chapter 5, where we used test-equating procedures and recent developments in test theory about subtest scores in classroom assessment. There is not much research that applied these methods in classroom assessments and as we showed they may contribute to better quality research in higher education.

From a theoretical perspective, the testing effect demonstrated by cognitive psychologists (e.g., Roediger & Karpicke, 2006), as well as assessment theory focusing on the feedback functionality (e.g., Hattie & Timperley, 2007) suggest that the learning process can be improved and that better student performance may be expected when practice tests are implemented. Chapter 4, however, showed that it is difficult to study whether this is actually the case, and that in practice student performance did not necessarily improve. This raises an important question for researchers in assessment: if the benefit of formative assessment is not visible or measurable in terms of student performance, how then should its success be evaluated and in what way should practitioners try to transfer the results of scientific research to their teaching practice?

Research in medical education has suggested that the implementation of cumulative or progressive testing may be beneficial (Kerdijk, Tio, Mulder, & Cohen-Schotanus, 2013; Kerdijk, Cohen-Schotanus, Mulder, Muntinghe, & Tio, 2015; Saint, Horton, Yool, & Elliot, 2015). In cumulative testing, students receive multiple exams throughout a given period in such a way that in each test, all the previously taught material is tested (including the material taught before the previous test). This ensures that students continue to study all the relevant material regularly, with each test counting towards the final grade. It is important to note, however, that this form of testing increases the summative function of a test (rather than remaining purely formative), which does reduce some of the benefits formative assessment can have on the learning process. Furthermore, Kerdijk et al (2015) did not find overall improved student performance in a group of students with cumulative tests, compared to a group of students with a final exam only. In addition, it is important to take the broader curriculum into account, as the research on cumulative assessment in medical education (particularly Kerdijk et al., 2013 and Kerdijk et al., 2015) is in the context of an integrated curriculum such that in each study period there are no separate courses. This in contrast to the curriculum of the social and behavioural sciences faculty in the present study, where students follow multiple separate courses at the same time, of which each is assessed separately. More research is necessary to determine whether cumulative testing is an ideal combination of formative and summative assessment functions to improve student learning in higher education.

As another example, several theories such as, active-learning (Prince, 2004), and student-engagement (Ashwin and McVitty 2015; Kahu 2013) all provide a convincing case for why flipping the classroom could be a good idea. Most research, however, attempted to demonstrate improved student performance (e.g., Davies, Dean and Ball 2013; Mason, Shuman and Cook 2013; McLaughlin et al. 2013; Pierce and Fox 2012; Street, Gilliland, McNeill and Royal 2015) rather than the intended behaviour change targeted by implementing the flipped classroom. By focusing on the targeted behaviour change it became evident in chapter 5 that students' self-regulation is also important to take into consideration. The importance and interaction of different theories in educational research will be much better understood when focusing on the targeted change of an educational innovation, rather than on the indirect outcome such as improved student performance. Thus chapter 5 provided an important contribution to the next steps in researching the implementation of the flipped classroom.

7.5 Contribution to practice

Based on the findings in chapter 2, an important practical finding was that the mode of exam administration did not influence the performance on exams consisting of multiple choice questions. However, when transition to computer-based exams also implies changing the type of question (e.g., using constructed response questions instead of multiple choice questions as in Dermo, 2009 and Peterson & Reider, 2002), care should be taken that the assessment give students a fair chance to demonstrate their mastery of the learning goals at the intended levels of knowledge. Given that students do not seem to prefer computer-based exams over paper-based exams, higher education institutes should carefully consider

the design and affordability of different computer-based exam applications. An application demonstrated by McNulty et al. (2007), for example, showed how test-taking strategies students were accustomed to in a paper-based exam could also be applied in computer-based exams. This could increase students control over the test-taking process and help reduce the stress students have during exams

The research conducted in chapter 4 also has several practical implications. When lecturers use incentives such as bonus points or extra credit so that all students benefit from formative assessment, this may have the unfortunate consequence that students do not use the assessment in a truly formative manner (as also shown by Kibble, 2007). On the other hand, keeping assessment truly formative, that is, when it is not part of the grade, means that students are required to be more in charge of their own learning process. This may result in students not using the provided formative assessment resources. There is no definitive way to solve this dilemma, and lecturers need to be aware of these issues when deciding how to implement formative assessment. Furthermore, it is important to take the whole course context into account. In the case of the two statistics courses, the course design already activated students with mandatory small-group practical meetings in addition to the large-scale lectures. In these courses the use of practice tests correlated weakly with student performance on the final exam. In the case of a course with voluntary large-scale lectures only the correlation between students' use of practice tests and the final exam score was much stronger.

When innovations are implemented in higher education with the intention to improve the learning process, the intended and potential benefits need to be communicated with students on a regular basis. Although students were informed of the potential benefits of the flipped classroom, the course evaluations showed that some students did not understand or believe in the potential of the flipped classroom to aid their learning and this might have influenced their decision to study as usual. As in chapter 4, the dilemma of whether to require students to participate in activities that are intended to facilitate learning was also evident in the case of the flipped classroom course in chapter 5.

For higher education institutes there is a difficult trade off: if student responsibility for their own learning is desired, then learning activities do not need to be mandatory. You can lead a horse to water, but you cannot make it drink. This is a particularly difficult principle to hold on to in the context of institutional accountability with performance incentives for students who perform well. Institutions need to reflect critically whether they are offering students all the means to be self-regulated learners, or whether the aim of self-regulation is an excuse not to invest in facilities that may support learning.

To conclude, a further discussion should take place between stakeholders concerning the expected output of educational innovations. If an innovation is implemented to improve the learning process of students, what exactly then is the expected outcome of this improvement and for whom this outcome is important? If the aim is to improve the learning process of students, then evidence should be collected on the learning process. A critical discussion should arise when the expected outcome is better student performance: Should a greater percentage of students pass at the first attempt? Should a greater percentage of students pass overall? Should the mean grade of a cohort increase? These

outcome measures could be informative to evaluate the effectiveness, but care should be taken to define a priori how big of an improvement is to be expected, and evidence concerning the mechanism by which this is achieved should also be collected (i.e. study behaviour, use of innovations) should be collected. With the collected evidence, the use of test-equating strategies, and/or taking into account the variation in grades could be fruitful avenues to then evaluate the results. By combining sources of information, and not exclusively focusing on the outcome of student performance, lecturers and universities may gain more insight in the effectiveness of assessment innovations. A continuing collaboration between research and practice is necessary to improve the quality of assessment and learning in higher education.

A

Appendices

Chapter 2

Table A2(a). The approach to taking computer-based exams and paper-based exams in general in cohort 2013/2014

2013/2014	Midterm exam M(SD)	Final exam M(SD)	*F*(1, 265)	*p*-value	Partial η^2
In this computer-based exam I was able to:					
a. Work in a structured manner	3.38 (1.16)	3.21 (1.13)	0.71	.40	.003
b. Monitor my progress	3.44 (1.20)	3.70 (1.15)	4.52	.03	.017
c. Concentrate well	2.98 (1.32)	3.52 (1.15)	14.94	<.001	.054
In paper-based exams in general I am able to:					
a. Work in a structured manner	4.19 (0.74)	4.25 (0.67)	0.59	.44	.002
b. Monitor my progress	4.08 (0.75)	4.17 (0.78)	1.36	.35	.005
c. Concentrate well	4.07 (0.74)	3.99 (0.79)	0.34	.56	.001

Table A2(b). The approach to taking computer-based exams and paper-based exams in general in cohort 2014/2015

2014/2015	Midterm exam M(SD)	Final exam M(SD)	*F*(1, 325)	*p*-value	Partial η^2
In this computer-based exam I was able to:					
a. Work in a structured manner	3.62 (1.06)	3.33 (1.17)	7.77	0.006	0.017
b. Monitor my progress	3.84 (1.10)	3.43 (1.17)	11.78	0.001	0.031
c. Concentrate well	3.79 (0.99)	3.42 (1.11)	11.39	0.001	0.031
In paper-based exams in general I am able to:					
a. Work in a structured manner	4.11 (0.79)	4.27 (0.66)	3.03	0.08	0.012
b. Monitor my progress	3.89 (0.89)	4.18 (0.77)	8.70	0.003	0.026
c. Concentrate well	3.95 (0.84)	4.09 (0.74)	2.61	0.11	0.008

Figure A4(a). Checking of model assumptions for Biopsychology

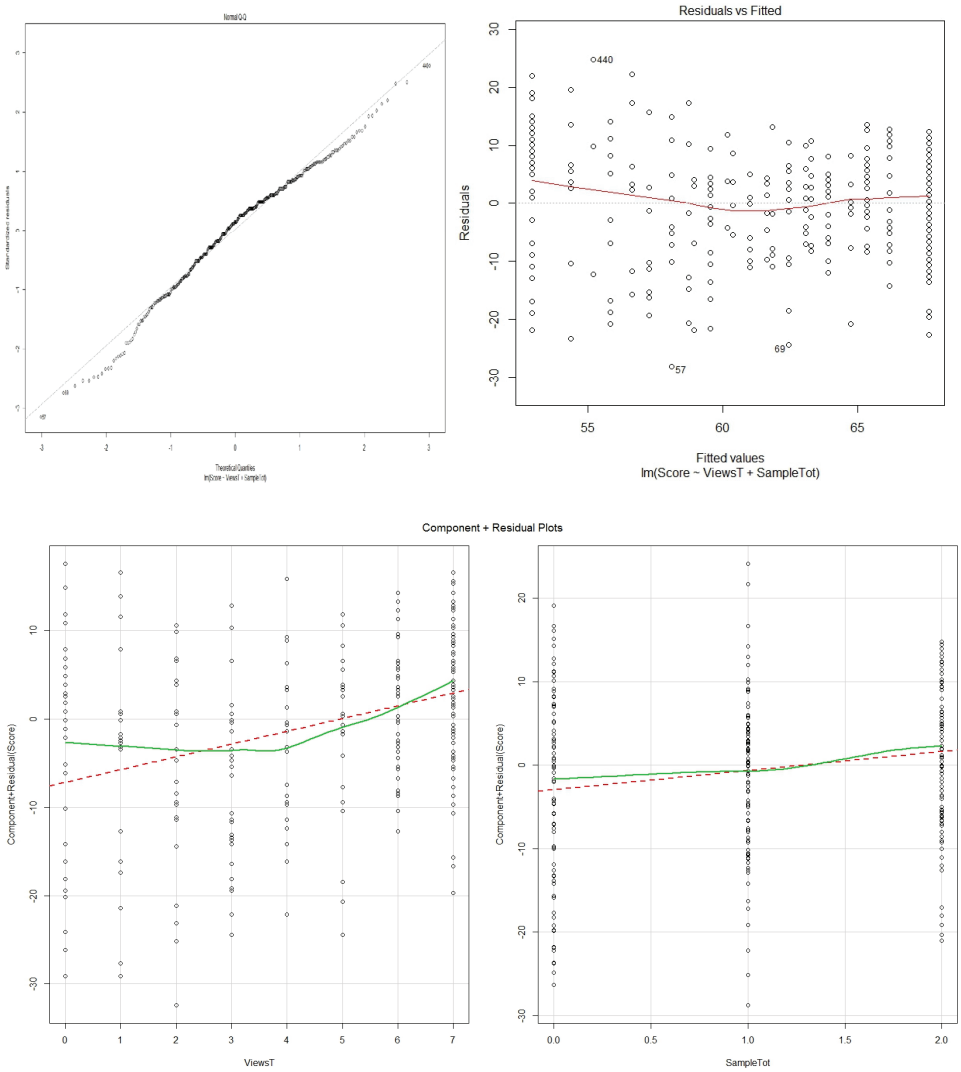


Figure A4(b). Checking of model assumptions for Statistics 1a

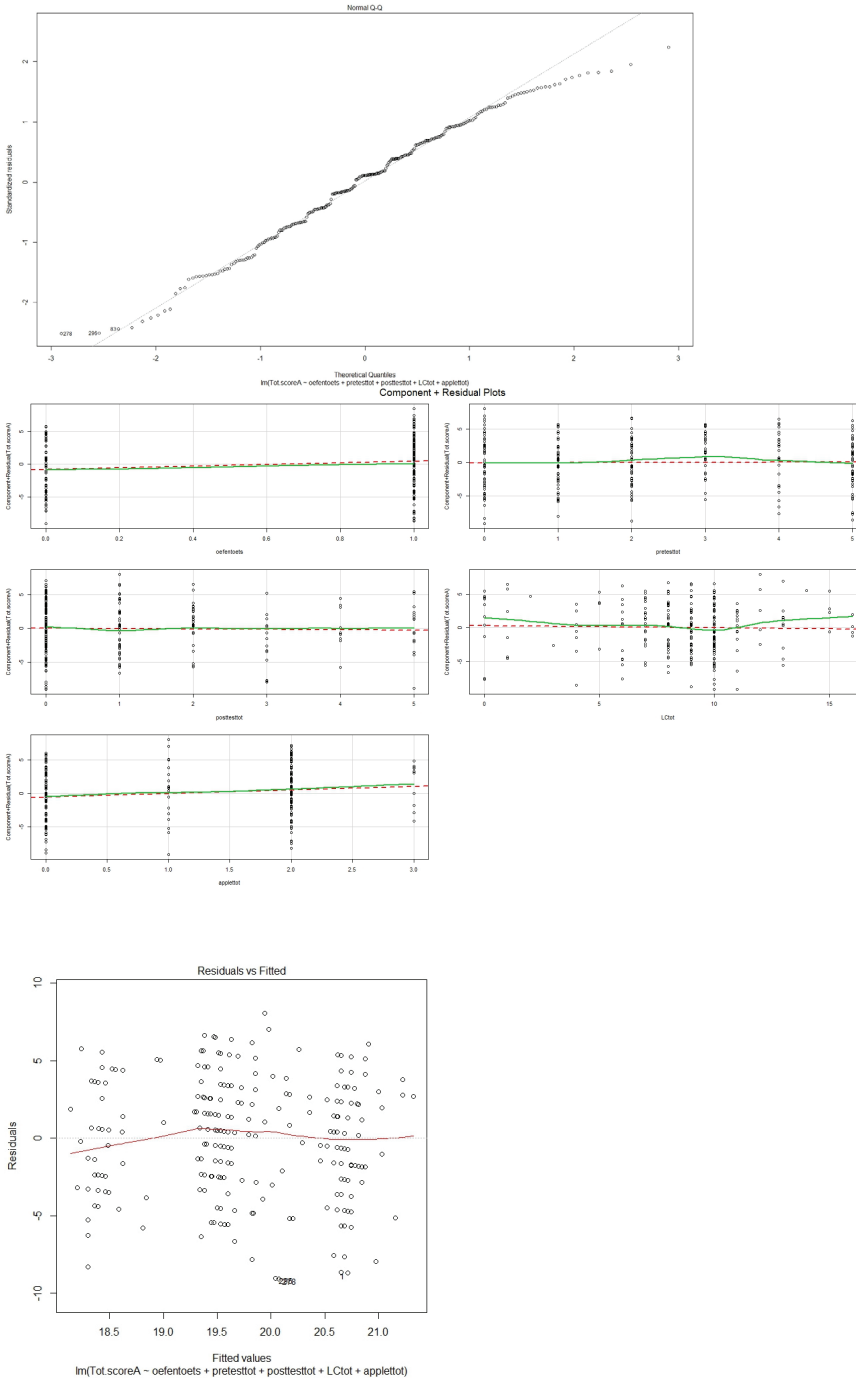
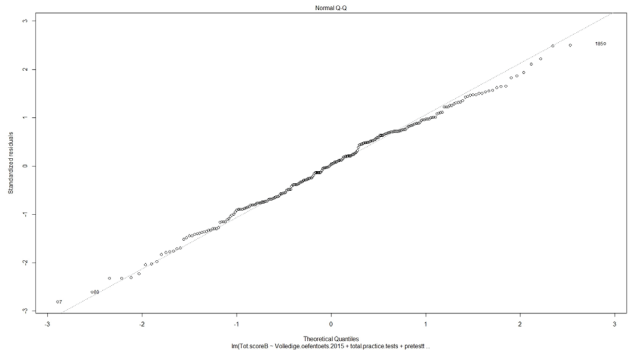
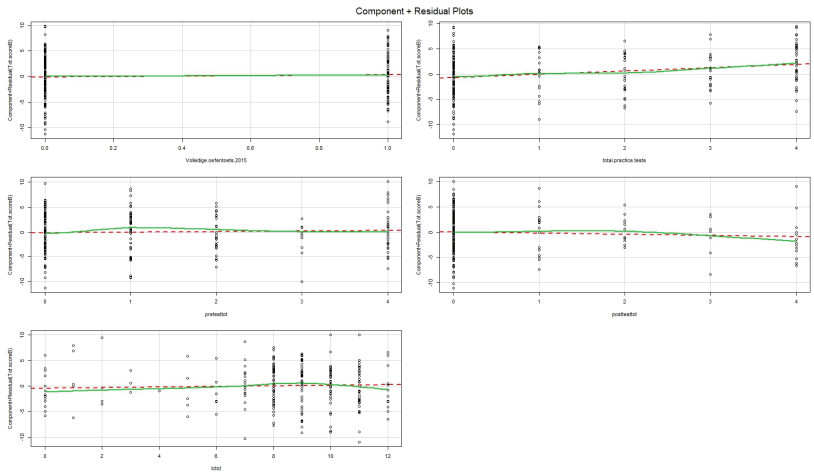
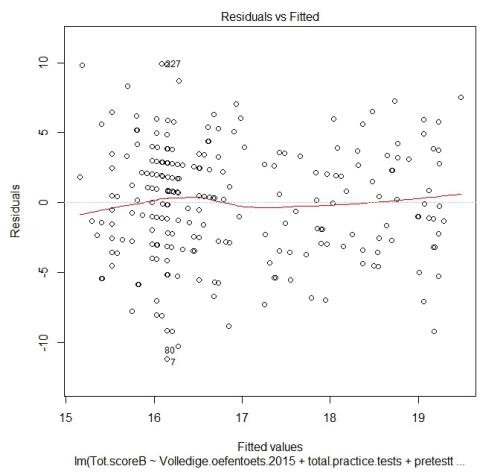


Figure A4(c). Checking of model assumptions for Statistics 1b







Chapter 5

Example of a study behaviour diary used for students in the flipped course. The original language used was Dutch. An English translation is given in *{italics}*. The software used for the online behaviour diaries was Qualtrics (www.qualtrics.com), and the option "export to word" was used to make the following document.

Deze vragenlijst gaat over je studietijdsbesteding voor statistiek 2 van afgelopen maandag tot donderdag. Zie nestor - cursusdocumenten voor meer informatie over dit onderzoek.
{This survey is about how you spent your time studying for statistics 2 from last Monday through Thursday. See course documents in Nestor for more information about this study}

Wanneer je van 10 maart tot 10 april minimaal 80% van de vragenlijsten invult maak je kans op één van de 20 kadobonnen ter waarde van 10 euro! 14 april worden de winnaars bekend gemaakt. *{If you fill out 80% or more of the surveys from 10th of March through 10th of April, you will be eligible to win one of the 20 gift vouchers each worth 10 euros. The winners will be informed on April 10}*

Voor vragen en opmerkingen over dit onderzoek kun je altijd bij mij terecht (zie onderstaande contact gegevens. *{For questions and comments about this study you can always contact: (contact details).}*

Meedoen aan dit onderzoek heeft geen enkele invloed op je resultaat voor dit vak. Deelname is geheel vrijwillig, dus je kunt op ieder moment stoppen. Wanneer je doorgaat naar de vragenlijst, geeft je automatisch toestemming voor deelname aan het onderzoek.Je kunt op ieder moment de gegevens die verkregen zijn uit dit onderzoek terugkrijgen, laten verwijderen uit de database, of laten vernietigen. *{Participating in this study has no influence on your final grade for the course. Participation is completely voluntary, and you may stop participating at any point in time. If you continue with the survey by pressing next, you consent to participating in this study. You may at any point in time contact the researcher and ask for your record to be destroyed or removed from the database.}*

Heel erg bedankt voor je medewerking! *{Thank you for your participation!}*

Q1

Wat is je s-nummer? *{What is your student number?}*

Dit is bedoeld om je antwoorden te kunnen koppelen aan je antwoorden op andere vragenlijsten in dit onderzoek. Je antwoorden worden echter wel anoniem en vertrouwelijk opgeslagen. *{This is necessary to be able to link your answers to the questions in this survey to the other surveys in this study. Your answers will be kept confidential an stored anonymously.}*

Q2

Heb je afgelopen maandag tijd besteed aan statistiek 2?

{Did you spend time on Statistics 2 last Monday?}

- ❍ Ja *{yes}*
- ❍ Nee *{no}*

If nee*{no}* Is Selected, Then Skip To **Q5**

Answer If **Q2** ja*{yes}* Is Selected:

Q3

Wat heb je maandag gedaan voor statistiek 2?

{What did you do for statistics 2 last Monday?}

- ❑ Flits college bekeken *{watched the video lecture}*
- ❑ studiestof gelezen *{read course material}*
- ❑ college sheets bekeken *{studied lecture slides}*
- ❑ Huiswerk voor deze week *{completed this week's homework}*
- ❑ samenvatting studiestof gemaakt *{summarized course material}*
- ❑ oefenvragen gemaak *{completed practice questions}*
- ❑ bijles gevolgd *{received tutoring}*
- ❑ anders *{other}* _____

If **Q3** Is Equal to 0, Then Skip To **Q5**

Options checked in **Q3** are carried forward to **Q4**

Q4 Hoeveel tijd heb je hier maandag aan besteed?

{How much time did you spend on this last Monday?}

For each option checked in **Q3** there is a drop down list with the following amounts available to select: 15 minutes, 30 minutes, 45 minutes, 1 hour, 1 hour 15 minutes, …, 5 hours

Q5 Heb je afgelopen dinsdag tijd besteed aan statistiek 2? (practicum aanwezigheid niet meegeteld)

{Did you spend time on Statistics 2 last Tuesday? (do not count practical attendance)}

- ❍ ja *{yes}*
- ❍ nee *{no}*

If nee*{no}* Is Selected, Then Skip To **Q8**

Answer If **Q5** ja *{yes}* Is Selected

Q6

Wat heb je dinsdag gedaan voor statistiek 2? (practicum aanwezigheid niet meegeteld)

{What did you do for statistics 2 on Tuesday? (not including practical attendance)}

- ❑ Flits college bekeken *{watched the video lecture}*
- ❑ studiestof gelezen *{read course material}*
- ❑ college sheets bekeken *{studied lecture slides}*
- ❑ Huiswerk voor deze week *{completed this week's homework}*
- ❑ samenvatting studiestof gemaakt *{summarized course material}*
- ❑ oefenvragen gemaak *{completed practice questions}*
- ❑ bijles gevolgd *{received tutoring}*
- ❑ anders *{other}* _____

If **Q6** Is Equal to 0, Then Skip To **Q8**

Options checked in **Q6** are carried forward to **Q7**

Q7

Hoeveel tijd heb je hier dinsdag aan besteed? *{How much time did you spend on these activities Tuesday?}*

For each option checked in **Q6** there is a drop down list with the following amounts available to select: 15 minutes, 30 minutes, 45 minutes, 1 hour, 1 hour 15 minutes, …, 5 hours

Q8

Heb je afgelopen woensdag tijd besteed aan statistiek 2? (practicum aanwezigheid niet meegeteld)

{Did you spend time on statistics 2 last Wednesday? (practical attendance not included)

- ❍ ja *{yes}*
- ❍ nee *{no}*

If nee*{no}* Is Selected, Then Skip **Q11**

Answer If **Q8** ja*{yes}* Is Selected

Q9

Wat heb je woensdag gedaan voor statistiek 2? *{What did you do for statistics 2 last Wednesday?}*

- ❑ Flits college bekeken *{watched the video lecture}*
- ❑ studiestof gelezen *{read course material}*
- ❑ college sheets bekeken *{studied lecture slides}*
- ❑ Huiswerk voor deze week *{completed this week's homework}*
- ❑ samenvatting studiestof gemaakt *{summarized course material}*
- ❑ oefenvragen gemaak *{completed practice questions}*
- ❑ bijles gevolgd *{received tutoring}*
- ❑ anders *{other}* _____

If **Q9** Is Equal to 0, Then Skip To **Q11**

Options checked in **Q9** are carried forward to **Q10**

Q10 Hoeveel tijd heb je hier woensdag aan besteed? *{How much time did you spend on these activities Wednesday?}*

For each option checked in **Q9** there is a drop down list with the following amounts available to select: 15 minutes, 30 minutes, 45 minutes, 1 hour, 1 hour 15 minutes, …, 5 hours

Q11

Heb je afgelopen donderdag tijd besteed aan statistiek 2? (practicum aanwezigheid niet meegeteld) *{Did you spend time on statistics 2 last Thursday? (practical attendance not included)}*

- ❍ ja *{yes}*
- ❍ nee *{no}*

If nee *{no}* Is Selected, Then Skip To **Q14**

Answer If **Q11** ja *{yes}* Is Selected

Q12

Wat heb je donderdag gedaan voor statistiek 2? *{What did you do for statistics 2 last Thursday?}*

- ❑ Flits college bekeken *{watched the video lecture}*
- ❑ studiestof gelezen *{read course material}*
- ❑ college sheets bekeken *{studied lecture slides}*
- ❑ Huiswerk voor deze week *{completed this week's homework}*
- ❑ samenvatting studiestof gemaakt *{summarized course material}*
- ❑ oefenvragen gemaak *{completed practice questions}*
- ❑ bijles gevolgd *{received tutoring}*
- ❑ anders *{other}* _____

If **Q12** Is Equal to 0, Then Skip To **Q14**

Options checked in **Q12** are carried forward to **Q13**

Q13 Hoeveel tijd heb je hier donderdag aan besteed? *{How much time did you spend on this last Thursday?}*

For each option checked in **Q12** there is a drop down list with the following amounts available to select: 15 minutes, 30 minutes, 45 minutes, 1 hour, 1 hour 15 minutes, …, 5 hours

Q14 Bedankt voor het invullen van de vragen! *{Thank you for completing this survey!}*
Tip: gebruik je agenda om bij te houden wat je elke dag voor statistiek 2 doet, zodat je minder moeite hoeft te doen om voor dagen die langer geleden zijn te onthouden wat je hebt gedaan en hoe lang je daar mee bezig bent geweest. *{Tip: use your diary to keep track of when and how long you study for statistics 2 so that it is less effort to remember when you studied as you fill out the survey.}*
Hieronder kun je opmerkingen of suggesties kwijt over deze vragenlijst.*{Below you can leave comments or suggestions about the survey}*

Chapter 6

Table A6(a). Number of assessment observations per faculty in each year

Faculty	2010	2011	2012	2013	2014	2015	Total
Theology	394	371	469	358	305	359	2,256
Law	7,683	7,391	7,281	4,703	5,051	4,844	36,953
Medicine	4,993	4,588	4,445	4,831	3,976	3,552	26,385
Science	10,377	10,735	10,991	12,053	11,818	12,235	68,209
Arts	11,923	11,021	10,686	11,325	10,477	10,366	65,798
Economy	15,514	14,461	13,495	14,392	13,954	12,136	83,952
Social	13,758	13,552	13,312	12,808	11,132	9,001	73,563
Philosophy	1,173	1,261	1,311	956	755	845	6,301
Spatial	2,337	2,108	1,935	1,890	2,034	1,372	11,676
Total	68,152	65,488	63,925	63,316	59,502	54,710	375,093

Table A6(b). Number of courses per faculty in each year

	2010	2011	2012	2013	2014	2015	Total
Theology	17	20	22	22	23	22	126
Law	28	28	28	26	20	17	147
Medicine	32	36	36	40	40	37	221
Science	101	104	106	101	106	104	622
Arts	208	206	171	168	174	167	1094
Economy	66	70	67	51	50	50	354
Social	69	73	74	75	69	67	427
Philosophy	21	20	19	24	13	13	110
Spatial	18	18	19	18	18	13	104
Total	560	575	542	525	513	490	3205

Table A6(c). Number of unique courses per faculty, and distribution of courses by number of cohorts included in the data

Faculty	1 cohort	2 cohorts	3 cohorts	4 cohorts	5 cohorts	6 cohorts	Total
Theology	7	4	3	3	6	10	33
Law	1	0	31	5	3	3	43
Medicine	22	20	7	5	2	18	74
Science	10	24	11	13	7	74	139
Arts	67	105	62	47	19	58	358
Economy	9	39	16	42	3	6	115
Social	8	32	6	25	9	32	112
Philosophy	8	6	8	9	0	5	36
Spatial	8	4	1	3	11	3	30
Total	140	234	145	152	60	209	940

Samenvatting

Inleiding

In dit proefschrift staat toetsing in het (universitair) hoger onderwijs centraal. De onderzoeken die zijn uitgevoerd in hoofdstuk twee tot en met zes zijn een verzameling van studies waarin de implementatie van verschillende innovaties op het gebied van toetsing in het onderwijs aan de Rijksuniversiteit Groningen zijn onderzocht.

Enkele recente belangrijke ontwikkelingen vormden de aanleiding voor het onderzoek in dit proefschrift. Deze zijn: de digitalisering van de maatschappij die ook in het onderwijs merkbaar is en tot verschillende veranderingen leidt; de groeiende studentaantallen, maar ook de politieke ontwikkelingen zoals het recent ingevoerde model van prestatiebekostiging voor hogeronderwijsinstellingen.

Doordat het aantal studenten in het hoger onderwijs is toegenomen in de afgelopen jaren (Hornsby & Osman, 2014), krijgen docenten te maken met steeds grotere groepen van soms wel honderden studenten. Hierdoor wordt de verhouding docent(en) ten opzichte van het aantal studenten erg klein, waardoor docenten maar zeer beperkte tijd en middelen ter beschikking hebben om de kwaliteit en voortgang van het leerproces van studenten te waarborgen.

De toenemende digitalisering is ook van belang in alle lagen van het onderwijs, dus ook voor het hoger onderwijs. Enerzijds is het niet meer mogelijk om toegang tot het hoger onderwijs te krijgen zonder de beschikking te hebben over digitale middelen, anderzijds blijft de rol van digitale middelen soms zeer beperkt in het onderwijs. Studenten moeten bijvoorbeeld dikwijls nog tentamens op papier maken, welke vervolgens ook met de hand worden nagekeken. En hoewel docenten gestimuleerd worden om digitale middelen in colleges te gebruiken, zien docenten hier soms van af en willen ze het gebruik van digitale middelen juist weer beperken Er ligt dus een uitdaging voor management en docenten om zo goed mogelijk gebruik te maken van digitale middelen op een manier die de kwaliteit van het leerproces ten goede komt.

Een andere aanleiding voor het onderzoek in dit proefschrift was de prestatiebekostiging van het hoger onderwijs in Nederland; in sommige is deze al eerder, op verschillende manieren ingevoerd (De Boer et al., 2015). De prestatiebekostiging in Nederland houdt in dat sinds 2012 afspraken zijn gemaakt tussen hogeronderwijsinstellingen– universitair en HBO – en de overheid. Wanneer instellingen de gestelde doelen halen binnen de afgesproken termijn, dan blijft de financiering van kracht. Wanneer een instelling niet de gestelde doelen behaald, wordt de financiering door de overheid gekort. Duidelijke indicatoren binnen deze afspraken zijn uitvalpercentages en het afstudeerrendement na vier jaar. Door deze prestatiebekostiging zou de kwaliteit van het onderwijs beter gewaarborgd worden. Hoewel minister Bussemaker (2014) heeft erkend dat kwaliteit niet alleen gemeten kan worden door uitval en rendement, bleven de andere kwaliteitsindicatoren vaag. Bussemaker (2014) stelde echter wel: “Nieuwe ontwikkelingen als open-online onderwijs bieden mogelijkheden om de kwaliteit van het hoger onderwijs verder te verbeteren. Het hoger onderwijs moet aan al deze

[2] Zie https://nos.nl/op3/artikel/2133931-laptops-in-de-collegezaal-streng-verboden.html

(bestaande en nieuwe) uitdagingen de juiste aandacht geven." In dit proefschrift wordt de implementatie van enkele van deze nieuwe mogelijkheden onderzocht in de praktijk.

Assessment en toetsen

In de wetenschappelijke literatuur wordt een onderscheid gemaakt tussen de termen 'assessment' en toetsen. Met assessment wordt het proces van informatieverzameling bedoeld waarmee men inzicht krijgt in de kennis en kunde van studenten. Deze informatie kan gebruikt worden voor diverse doeleinden, zoals de richting van vervolginstructie, diagnose van sterke/zwakke kanten van studenten en/of het nemen van beslissingen over studenten. Een toets daarentegen is een verzameling van vragen of opdrachten die tot doel heeft de kennis en/of kunde op een specifiek tijdstip van een student te meten. Een toets kan en is meestal onderdeel van een assessment-programma. In dit proefschrift gaan sommige hoofdstukken specifiek over toetsen (de hoofdstukken 2, 3 en 6), terwijl andere hoofdstukken over assessment gaan (de hoofdstukken 4 en 5).

Toetsen kunnen met verschillende bedoelingen ontworpen en/of ingezet worden. Wanneer het vooral belangrijk is om een beslissing te nemen – of om te selecteren – wordt de functie van een toets omschreven als 'summatief'. Voorbeelden van summatieve toetsen zijn toelatingstoetsen of tentamens aan het einde van een onderwijsperiode. Formatieve toetsen daarentegen zijn bedoeld om het leerproces te bevorderen en/of bij te sturen. Dit onderscheid wordt ook wel omschreven als het 'toetsen van leren' of het 'toetsen voor leren' (Schuwirth & van der Vleuten, 2011). In de praktijk is het onderscheid tussen formatieve en summatieve toetsen niet altijd even duidelijk, zoals duidelijk wordt in hoofdstukken 3 tot en met 5 van dit proefschrift.

Er zijn verschillende tradities van onderzoek naar de kwaliteit van toetsing. Men spreekt in de literatuur van "high-stakes" grootschalige toetsing, wanneer beslissingen op basis van de toetsen van groot belang zijn en voor zeer grote aantallen worden uitgevoerd. De bekendste vorm van grootschalig high-stakes toetsing in Nederland is de eindtoets van de basisschool die mede bepalend is naar welk vervolgonderwijs kinderen gaan. Ook in andere landen zijn er dergelijke "high-stakes" toetsen zoals bijvoorbeeld de SAT in de Verenigde Staten, die gebruikt wordt om studenten te selecteren voor het hoger onderwijs. Deze toetsen zijn regelmatig in het nieuws vanwege de vermeende negatieve effecten op de selectie van minderheden in het hoger onderwijs. Wat minder bekend bij dit soort grootschalige toetsen, is dat deze toetsen – los van het gebruik in de praktijk – aan strenge kwaliteitseisen moeten voldoen, waardoor hier doorgaans veel onderzoek naar gedaan wordt. Dit heeft bijgedragen aan de ontwikkeling van theorie en statistische methoden om de toetsen te beoordelen.

Naast de onderzoekstraditie van het grootschalige toetsen is er recent meer aandacht voor het kleinschaliger "classroom testing" waarbij toetsing vooral in dienst staat van het onderwijsleerproces. Deze toetsen worden doorgaans ontwikkeld door docenten voor relatief kleine groepen leerlingen of studenten. Vaak zijn niet de tijd en middelen beschikbaar om uitgebreid de kwaliteit van de toetsen te onderzoeken zoals dit wel het geval is bij grootschalige "high-stakes" toetsen. In de wetenschappelijke

literatuur is een groeiende interesse om het onderzoek naar grootschalig en kleinschalig toetsen te integreren. In dit proefschrift worden ook methoden die ontwikkeld zijn voor grootschalige toetsing toegepast bij kleinschalig “classroom testing” in het universitair onderwijs.

De context van het onderzoek in dit proefschrift

De studies die in dit proefschrift staan beschreven zijn uitgevoerd aan de Rijksuniversiteit Groningen (RUG) en de meeste daarvan zijn gedaan in het propedeusejaar. De resultaten behaald in het eerste jaar van de bachelor zijn belangrijk omdat er een sterk verband is tussen deze resultaten en de resultaten in het vervolg van de opleiding (Niessen, Meijer & Tendeiro, 2016) en omdat studenten te maken hebben met het bindend studieadvies (BSA). Het BSA is gebaseerd op een minimaal aantal studiepunten dat een student moet behalen – aan de RUG bijvoorbeeld 45 van de 60 ECTS – in het eerste jaar om door te mogen gaan met hun studie. Dus wanneer een student dit aantal niet haalt, mogen ze de opleiding niet meer vervolgen. Hier zijn ook financiële implicaties mee gemoeid: wanneer een student na februari besluit te stoppen – of het BSA niet haalt – dan blijft het collegegeld verschuldigd. Hierdoor zijn alle studieresultaten in het eerste jaar in de beleving van studenten ook wel te omschrijven als “high stakes”.

Samenvatting van de resultaten van het uitgevoerde onderzoek

De studies in dit proefschrift zijn uitgevoerd in samenwerking met docenten die hun onderwijs wilden verbeteren door veranderingen in toetsing door te voeren, waarbij veelal digitale middelen werden ingezet. In hoofdstuk 2 tot en met 6 worden de verschillende studies beschreven, Hoofdstuk 1 bevat een algemene introductie en in hoofdstuk 7 wordt op de resultaten teruggeblikt. Hieronder volgt een korte samenvatting van elk onderzoek:

In hoofdstuk 2 wordt de implementatie van digitale tentamens onderzocht. Digitale tentamens bieden mogelijkheden om de kwaliteit van toetsen te verbeteren en het toetsproces te vergemakkelijken. Daarentegen is het van belang om ervoor te zorgen dat de prestaties op de traditionele en digitale toetsen vergelijkbaar zijn, dat de studenten de toetsen als eerlijk ervaren en dat de stress die deze nieuwe vorm van toetsing met zich meebrengt minimaal is (Whitelock, 2009). De belangrijkste onderzoeksvragen in het onderzoek dat wordt beschreven in hoofdstuk 2 waren: is er een verschil in de prestatie tussen studenten die digitaal en schriftelijke werden getoetst? En: hoe ervaren studenten digitale toetsen? Dit is onderzocht door middel van een zogeheten ‘veldexperiment’, een experiment buiten een laboratorium. Studenten die het vak ‘Biopsychologie’ volgden – in 2013/2014 in de Nederlandstalige psychologieopleiding en in 2014/2015 in de Engelstalige psychologieopleiding – maakten op willekeurige basis ofwel de eerste deeltoets ofwel de tweede deeltoets digitaal en de ander op papier. De toetsresultaten van de groep studenten die de deeltoets op papier had gemaakt waren nagenoeg hetzelfde vergeleken met de groep studenten die de deeltoets digitaal had gemaakt, voor zowel de eerste als tweede deeltoets. Ongeveer een kwart van de studenten bleek een voorkeur te hebben voor de digitale toets, ongeveer de helft een voorkeur voor

de papieren versie en ongeveer een kwart had geen voorkeur (met uitzondering van de tweede deeltoets van de internationale opleiding waarbij de voorkeur voor digitale en papieren tentamens ongeveer gelijk was). Aangezien ook uit de literatuur blijkt dat studenten geen grote voorkeur hebben voor digitale toetsen, is het belangrijk dat instellingen de overgang naar digitaal toetsen zo inrichten dat studenten controle ervaren over het tentamen. Uit aanvullend kwalitatief onderzoek bleek dat dit op verschillende manieren zou kunnen worden gerealiseerd: door studenten bijvoorbeeld de mogelijkheid te bieden om bij digitale toetsen te kunnen onderstrepen, doorhalingen en markeringen te maken of door het flexibeler aanbieden van vragen.

In hoofdstuk 3 wordt het nut van het rapporteren van deelscores op een toets onderzocht. Gegeven de beperkte tijd en middelen van docenten in het hoger onderwijs, zou het wenselijk kunnen zijn om op efficiënte wijze diagnostische feedback aan studenten te geven door middel van het rapporteren van deelscores. Dit gebeurt soms bij “large-scale testing” en is ook steeds vaker een afweging bij “classroom testing”. Er is echter ook wetenschappelijke literatuur die aantoont dat dit maar beperkt zinvol is (bijv. Sinharay, 2010). In hoofdstuk 3 wordt aan de hand van twee verschillende tentamens geïllustreerd hoe onderzocht kan worden of het zinvol is om naast de totaalscore op de hele toets ook deelscores te rapporteren. Voor een van de tentamens werd naast de totaalscore ook deelscores van verschillende soorten kennis berekend. Voor het andere tentamen werd een deelscore berekend voor de open vragen en voor de meerkeuze vragen. In beide gevallen bleek het niet zinvol om de deelscores te rapporteren, dit kwam doordat deelscores relatief onbetrouwbaar waren en hoog correleerden met de totaalscore. Wel was interessant dat een deel van de open vragen sterk bijdroegen aan het vergroten van de meetprecisie van het totale tentamen.

In hoofdstuk 4 en 5 wordt onderzocht wat de effecten waren van het gebruik van digitale leermiddelen om het leerproces en de resultaten te verbeteren. In hoofdstuk 4 is de implementatie van digitale oefentoetsen onderzocht in verschillende vakken. In het eerste deel van deze studie is gekeken naar het gebruik van oefentoetsen door studenten in twee statistiekvakken en een vak over biopsychologie. Voor de statistiekvakken bleek dat het gebruik van oefentoetsen nauwelijks voorspellend was voor het tentamencijfer van de studenten. Voor het vak over biopsychologie daarentegen was er een duidelijk positieve samenhang tussen de mate van gebruik van oefentoetsen en de tentamenscore. Een mogelijke reden voor de verschillende resultaten kan de cursusinrichting zijn – de studenten bij de statistiekvakken hadden naast hoorcolleges en oefentoetsen ook verplichte werkgroepen en huiswerk. Bij biopsychologie waren geen verplichtingen behalve het maken van het tentamen en konden studenten facultatief naar de hoorcolleges. Tevens is het belangrijk om te erkennen dat een positieve samenhang tussen oefentoets gebruik en tentamenscore niet noodzakelijk iets zegt over de effectiviteit van de oefentoetsen. Een alternatieve verklaring kan zijn dat vooral de gemotiveerde studenten gebruik hebben gemaakt van de oefentoetsen en dat deze studenten ook een goed cijfer hadden behaald wanneer er geen oefentoetsen beschikbaar waren. Daarom is ook een tweede deelstudie uitgevoerd voor het vak biopsychologie.

In de tweede studie van hoofdstuk 4 is gekeken naar het verschil in gemiddelde toetsscores van een cohort biopsychologiestudenten die geen beschikking hadden over oefentoetsen en twee cohorten biopsychologiestudenten die gedeeltelijk of geheel de beschikking hadden over oefentoetsen. Er werd gebruik gemaakt van test-equating om de scores van de verschillende cohorten met elkaar te vergelijken. Uit de resultaten bleek dat er nauwelijks verschil was in de prestaties tussen de groepen die wel of geen beschikking hadden over de oefentoetsen. Het onderzoek in dit hoofdstuk liet ook zien dat het in de praktijk soms lastig is om bevindingen uit (experimenteel) wetenschappelijk onderzoek daadwerkelijk in de praktijk te implementeren. In de praktijk is het belangrijkste dilemma bij het invoeren van formatief toetsen de mate waarin het verplicht zou moeten worden. Hierbij is het belangrijk te beseffen dat wanneer deelname aan formatieve toetsing verplicht wordt, de toets automatisch summatiever wordt. Eerder onderzoek heeft aangetoond dat zelfs niet-dwingende maatregelen zoals bonuspunten negatieve gevolgen kunnen hebben voor de formatieve werking van de toets (Kibble, 2007). Aan de andere kant betekent dat het dat bij daadwerkelijk formatieve toetsing studenten de verantwoordelijkheid moeten nemen voor hun eigen leerproces, met het risico dat de docent middelen beschikbaar stelt die niet gebruikt worden.

In hoofdstuk 5 staat onderzoek naar de implementatie van de “flipped classroom” centraal. In de flipped classroom gaan studenten tijdens het hoorcollege actief aan de slag met de leerstof en vind kennisoverdracht – met hulp van bijvoorbeeld videoclips – ook plaats voorafgaand aan de les (Abeysekera & Dawson, 2015; Street, Gilliland, McNeil, & Royal, 2015). De populariteit van de flipped classroom neemt toe in het (universitair) hoger onderwijs. Hoewel er enig onderzoek is gedaan naar de prestaties van groepen studenten die flipped classroom onderwijs volgden, is er nog weinig bekend over het studiegedrag van studenten in de flipped classroom. Dit studiegedrag is belangrijk omdat het centraal staat in het leerproces en prestaties van de flipped classroom. In hoofdstuk 5 is een studie verricht waarbij een groep studenten een statistiekvak volgden in de vorm van de flipped classroom en een groep studenten die een statistiekvak volgden in de traditionele vorm. Twee keer in de week werden studenten gevraagd om in te vullen hoeveel tijd zij hadden besteed aan het vak en welke studieactiviteiten ze hadden ondernomen voor het statistiekvak. Uit de resultaten bleek dat het studiepatroon van de twee groepen gedurende het vak sterk op elkaar leek en tevens dat de gemeten studiegedrag (tijd en activiteiten) ook niet een sterke samenhang vertoonde met de verkregen tentamenresultaten. In een verdere exploratie van de vakevaluaties voor de groep flipped classroom-studenten werd specifiek gekeken naar evaluaties met betrekking tot hun studiegedrag en perceptie van de “flipped classroom”. Sommige studenten vonden inderdaad dat de “flipped classroom” hun leerproces ondersteunde. Andere studenten daarentegen waren om diverse redenen niet bereid om hun studiegedrag te veranderen in de flipped classroom. Dit biedt interessante mogelijkheden voor vervolg onderzoek. Hoewel er voldoende wetenschappelijk theoretische gronden zijn die zouden moeten ondersteunen dat de flipped classroom een goed idee is, is veel onderzoek gericht op het aantonen van verbeterde prestaties en niet op het gedrag van studenten dat ten grondslag ligt aan de theorie en implementatie

van de “flipped classroom”. Het onderzoek in dit hoofdstuk liet zien dat de zelfregulatie van studenten en hun bereidheid om mee te gaan met de gedragsverandering belangrijk is bij de implementatie van de flipped classroom.

Geïnspireerd door het onderzoek in hoofdstukken 2 tot en met 5, heeft het onderzoek in hoofdstuk 6 een meer methodologisch karakter. In onderzoek naar innovaties of veranderingen in het onderwijs is de prestatie van studenten dikwijls de belangrijkste uitkomst. In de meest gangbare type onderzoek worden of bestaande groepen studenten over verschillende jaren met elkaar vergeleken, of worden twee bestaande verschillende groepen in dezelfde periode met elkaar vergeleken. Dit type onderzoek wordt gebruikt omdat het meestal onmogelijk is om willekeurig studenten aan groepen toe te wijzen zoals bijvoorbeeld gebeurt bij gerandomiseerde gecontroleerde trials. Het nadeel van het gebruik van bestaande groepen studenten is dat alternatieve variabelen naast “de treatment” van invloed kunnen zijn. Dus de verschillen in de prestaties van groepen studenten in verschillende condities hoeven niet het resultaat van bijvoorbeeld ingevoerde onderwijsvernieuwingen. De onderzoeksvragen in hoofdstuk 6 was: in welke mate fluctueren de prestaties van studenten in eerstejaarsvakken over tijd en tussen vakken? Hoe kan deze informatie gebruikt worden om onderwijsinnovaties te evalueren? Om deze vraag te beantwoorden zijn de resultaten van eerstejaars vakken over een periode van zes jaar aan de Rijksuniversiteit Groningen geanalyseerd. In totaal kon 17% van de variatie in cijfers toegekend worden aan fluctuatie over tijd en tussen vakken. Verder kon 40% van de variatie in slagingspercentages worden toegekend aan fluctuaties over tijd en tussen vakken. Gebruikmakend van deze informatie wordt in hoofdstuk 6 geïllustreerd wanneer verschillen in gemiddelde cijfers tussen groepen binnen de natuurlijk te verwachten fluctuatie valt en wanneer er sprake is van een betekenisvol verschil.

Beperkingen van dit onderzoek en toekomstig onderzoek

In het onderzoek in dit proefschrift hebben we geprobeerd een bijdrage te leveren aan een antwoord op de verschillende assessment vragen in het hoger onderwijs. Dit onderzoek werd in de praktijk uitgevoerd, hetgeen ook een aantal beperkingen met zich meebracht. Een beperking van het onderzoek in dit proefschrift is dat het plaats vond aan een enkele universiteit in Nederland. Hierdoor is het mogelijk dat de resultaten niet direct te generaliseren zijn naar andere onderwijsprogramma’s, naar het hbo of instellingen in andere landen.

Een andere beperking, zoals hierboven besproken, is dat het vaak niet mogelijk wanneer onderwijsvernieuwingen worden ingevoerd om experimentele designs te gebruiken, waardoor het moeilijk is om het causale effect van een implementatie vast te stellen. Onderzoek op grote schaal, in de vorm van veldexperimenten, zou misschien een goede aanpak kunnen zijn in de toekomst. Hier zouden verschillende opleidingen binnen meerdere instellingen kunnen deelnemen, zodat wellicht cursussen willekeurig kunnen worden toebedeeld aan onderwijsvernieuwingen.

Een andere belangrijke vraag die nader onderzocht kan worden is de vraag naar de relatie tussen onderwijsvernieuwingen en de context van de vakken waarin deze

geïmplementeerd worden. Het was, bijvoorbeeld, opvallend dat in de statistiekvakken die onderzocht zijn, zowel in hoofdstuk 4 als hoofdstuk 5, dat er nauwelijks een relatie werd gevonden tussen studiegedrag, gebruik van oefentoetsen en het eindcijfer. De statistiekvakken werden naast de vrijblijvende hoorcolleges doorgaans gekenmerkt door meerdere werkvormen en verplichtingen zoals werkgroepen en huiswerk. De onderwijsvernieuwingen zoals de "flipped classroom" en het aanbieden van oefentoetsen hadden misschien geen meerwaarde ten opzichte van de bestaande "good practices". Grootschaliger onderzoek is van belang, om zowel contextuele factoren beter in kaart te brengen, als ook omdat er veel kleinschalige studies in specifieke onderwijscontexten worden gepubliceerd. Dit kan voor vertekening zorgen gezien vooral positieve en statistisch significante resultaten worden gepubliceerd.

Tot slot is het belangrijk dat er wordt nagedacht door diverse belanghebbenden over de te verwachten uitkomsten van onderwijsinnovaties. Wanneer onderwijsinnovaties ten doel hebben om het leerproces van studenten te verbeteren, wat is dan precies de verwachtte uitkomst en voor wie is deze belangrijk? Als het doel is het leerproces te verbeteren, dan moet ook daadwerkelijk evidentie worden verzameld op dit gebied. Ook zou kunnen worden gekeken wat onder "verbeterde prestaties" wordt verstaan. Betekent dit een groter slagingspercentage bij het eerste tentamen of na meerdere tentamens? Voor de gehele groep of voor de minder goede studenten? Deze uitkomsten kunnen informatief zijn voor de effectiviteit van innovaties, maar het is belangrijk om van tevoren te definiëren hoe groot een te verwachten verbetering zou mogen zijn en om informatie over het mechanisme dat tot de verandering leidt te verzamelen. Hierbij is het ook belangrijk om analyses te gebruiken die rekening houden met de praktijk, zoals natuurlijke schommelingen in cijfers en verschillen in de moeilijkheid van tentamens. Door verschillende informatiebronnen te gebruiken en niet exclusief op uitkomsten te focussen, kunnen docenten en onderzoekers beter inzicht krijgen in wanneer onderwijsinnovaties – waaronder innovaties gericht op toetsing – effectief zijn in de praktijk. Een samenwerking is nodig tussen onderzoek en onderwijspraktijk om de kwaliteit van het leren en toetsen in het hoger onderwijs te verbeteren.

R

References

Abeysekera, L., & Dawson, P. (2015). Motivation and cognitive load in the flipped classroom: definition, rationale and a call for research. *Higher Education Research and Development*, *34*(1), 1-14. doi:10.1080/07294360.2014.934336

Alexander, F. K. (2000). The changing face of accountability: Monitoring and assessing institutional performance in higher education. *The Journal of Higher Education, 71*, 411-431. doi:10.2307/2649146

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (2014). *Standards for Educational and Psychological Testing.* Washington: American Educational Research Association.

Anakwe B. (2008). Comparison of student performance in paper-based versus computer-based testing. *Journal of Education for Business, 84*(1), 13-17. doi: 10.3200/JOEB.84.1.13-17

Angus, S. D., & Watson, J. (2009). Does regular online testing enhance student learning in the numerical sciences? Robust evidence from a large data set. *British Journal of Educational Technology, 40*(2), 255-272. doi:10.1111/j.1467-8535.2008.00916.x

Aojula, H., Barber, J., Cullen, R., & Andrews, J. (2006). Computer-based, online summative assessment in undergraduate pharmacy teaching: The Manchester experience. *Pharmacy Education, 6*(4), 229–236. doi:10.1080/15602210600886209

Apostolou, B., Blue, M. A., & Daigle, R. J. (2009). Student perceptions about computerized testing in introductory managerial accounting. *Journal of Accounting Education, 27*(2), 59-70. doi:10.1016/j.jaccedu.2010.02.003

Ashwin, P., & McVitty, D. (2015). The meanings of student engagement: implications for policies and practices. In The European Higher Education Area (pp. 343-359). Springer International Publishing. doi:10.1007/978-3-319-20877-0_23

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015) Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1-48. doi:10.18637/jss.v067.i01

Battauz, M. (2015). equateIRT: An R package for IRT test equating. *Journal of Statistical Software, 68*(7), 1-22. doi:10.18637/jss.v068.i07

Bayazit A., & Askar, P. (2012). Performance and duration differences between online and paper-pencil tests. *Asia Pacific Educational Review, 13*(2), 219-226. doi: 10.1007/s12564-011-9190-9

Beatty, A. S., Walmsley, P. T., Sackett, P. R., & Kuncel, N. R. (2015). The reliability of college grades. *Educational Measurement: Issues and Practice, 34*(4)*,* 31-40. doi:10.1111/emip.12096

Biggs, J., Kember, D., & Leung, D. Y. (2001). The revised two-factor study process questionnaire: R-SPQ-2F. *British Journal of Educational Psychology*, *71*(1), 133-149. doi:10.1348/000709901158433

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5*(1), 7-74. doi:10.1080/0969595980050102

Black, P., & Wiliam, D. (2003). 'In praise of educational research': Formative assessment. *British Educational Research Journal, 29*(5), 623-637. doi:10.1080/0141192032000133721

Boevé, A. J., Meijer, R. R., Albers, C. J., Beetsma, Y., & Bosker, R. J. (2015). Introducing computer-based testing in high-stakes exams in higher education: Results of a field experiment. *PloS one, 10*(12), doi:10.1371/journal.pone.0143616

Brookhart, S. M., Guskey, T. R., Bowers, A. J., McMillan, J. H., Smith, J. K., Smith, L. F., ... & Welsh, M. E. (2016). A century of grading research: Meaning and value in the most common educational measure. *Review of Educational Research, 86*(4), 803-848. doi:10.3102/0034654316672069

Bussemaker, J. (2014). Brief aan de voorzitter van de Tweede kamer der Staten Generaal: Voortgang Hoofdlijnenakkoorden en Prestatieafspraken Hoger Onderwijs en Onderzoek. The Hague, Netherlands: Ministerie van Onderwijs, Cultuur en Wetenschap

Burnard, P., Gill, P., Stewart, K., Treasure, E., & Chadwick, B. (2008). Analysing and presenting qualitative data. *British Dental Journal*, *204*(8), 429-432. doi:10.1038/sj.bdj.2008.292

Cagiltay, N., & Ozalp-Yaman, S. (2013). How can we get benefits of computer-based testing in engineering education?. *Computer Applications in Engineering Education 21*(2), 287-293. doi:10.1002/cae.20470

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). IRTPRO for windows [computer software]/ Scientific Software International, Lincolnwood, IL

Cantillon, P., Irish, B., & Sales, D. (2004). Using computers for assessment in medicine. *BMJ, 329*, 606-609. doi:10.1136/bmj.329.7466.606

Carey, T. A., & Stiles, W. B. (2015). Some problems with randomized controlled trials and some viable alternatives. *Clinical Psychology & Psychotherapy, 23*(1), 87-95. doi:10.1002/cpp.1942

Carpenter, S. K., Rahman, S., Lund, T. J., Armstrong, P. I., Lamm, M. H., Reason, R. D., & Coffman, C. R. (2017). Students' Use of Optional Online Reviews and Its Relationship to Summative Assessment Outcomes in Introductory Biology. *CBE-Life Sciences Education, 16*(2), ar23. doi:10.1187/cbe.16-06-0205

Carrillo-de-la-Peña, M. T., Bailles, E., Caseras, X., Martínez, À., Ortet, G., & Pérez, J. (2009). Formative assessment and academic achievement in pre-graduate students of health sciences. *Advances in Health Sciences Education, 14*(1), 61-67. doi:10.1007/s10459-007-9086-y

Cizek, C. K. (2009). Reliability and validity of information about student achievement: Comparing large scale and classroom testing effects. *Theory into Practice, 48*(1), 63-71. doi:10.180/004058408408025.77627

Clariana, R, & Wallace, P. (2002). Paper-based versus computer-based assessment: key factors associated with the test mode effect. *British Journal of Educational Technology, 33*(5), 593-602. doi:10.1111/1467-8535.00294

Credé, M., & Phillips, L. A. (2011). A meta-analytic review of the Motivated Strategies for Learning Questionnaire*. Learning and Individual Differences, 21*(4), 337-346. doi:10.1016/j.lindif.2011.03.002

Credé, M., Roch, S. G., & Kieszczynka, U. (2010). Class attendance in college: A meta-analytic review of the relationship of class attendance with grades and student characteristics. *Review of Educational Research, 80*(2), 272-295. doi:10.3102/0034654310362998

Cronbach, L. J. (1977). Essentials of Psychological Testing. New York, Harper & Brothers.

Csapo, B., Ainley, J., Bennett, R. E., Latour, T., & Law N. (2012). Technological issues for computer-based assessment. Griffin P., McGaw B., Care E. (eds). Assessment and teaching of 21st century skills. Dordrecht, Netherlands: Springer,. pp. 143

Davies, R. S., Dean, D. L., & Ball, N. (2013). Flipping the classroom and instructional technology integration in a college-level information systems spreadsheet course. *Educational Technology Research and Development, 61*(4), 563-580. doi:10.1007/s11423-013-9305-6

De Boer, H. F., Jongbloed, B. W. A., Benneworth, P. S., Cremonini, L., Kolster, R., Kottmann, A., … & Vossensteyn, J. J. (2015). Performance-based funding and performance agreements in fourteen higher education systems (Report for the Ministry of Education, Culture, and Science). Enschede, Netherlands: Center for Higher Education Policy Studies.

De Kleijn, R. A., Bouwmeester, R. A., Ritzen, M. M., Ramaekers, S. P., & Van Rijen, H. V. (2013). Students' motives for using online formative assessments when preparing for summative assessments. *Medical Lecturer, 35*(12), e1644-1650. doi:0.3109/0142159X.2013.826794

Deci, E. L., Vallerand, R. J., Pelletier, L. G., & Ryan, R. M. (1991). Motivation and education: The self-determination perspective. *Educational Psychologist, 26*(3-4), 325-246. doi:10.1080/00461520.1991.9653137

Dermo, J. (2009). e-Assessment and the student learning experience: A survey of student perceptions of e-assessment. *British Journal of Educational Technology, 40*(2), 203-214. doi:10.1111/j.1467-8535.2008.00915.x

Deutsch, T., Herrmann, K., Frese, T., & Sandholzer, H. (2012). Implementing computer-based assessment–a web-based mock examination changes attitudes. *Computers & Education, 58*(4), 1068-1075. doi:10.1016/j.compedu.2011.11.013

Dobson, J. L. (2008). The use of formative online quizzes to enhance class preparation and scores on summative exams. *Advances in Physiology Education, 32*(4), 297-302. doi:10.1152/advan.90162.2008

Dollinger, S. J., Matyja, A. M., & Huber, J. L. (2008). Which factors best account for academic success: Those which college students can control or those they cannot? *Journal of Research in Personality, 42*(4), 872-885. doi:10.1016/j.jrp.2007.11.007

Dove, A. (2013). Students' perceptions of learning in a flipped statistics class. In R. McBride & M. Searson (Eds.), Proceedings of Society for Information Technology & Lecturer Education International Conference 2013 (pp. 393-398). Chesapeake, VA: Association for the Advancement of Computing in Education (AACE).

Draper, N. R., & Smith, H. (2014). Applied regression analysis. New York: John Wiley & Sons.

El Shallaly, G. E., & Mekki, A. M. (2012). Use of computer-based clinical examination to assess medical students in surgery. *Educational Health, 25*(3), 148-152. doi: 10.4103/1357-6283.109789

Elo, S., & Kyngäs, H. (2008). The qualitative content analysis process. *Journal of Advanced Nursing, 62*(1), 107-115. doi:10.1111/j.1365-2648.2007.04569.x

Embretson, S. E. & Reise, S.P. (2000). Item response theory for psychologists. Mahwah NJ: Erlbaum.

Escudier, M. P., Newton, T. J., Cox, M. J., Reynolds, P. A., & Odell, E. W. (2011). University students' attainment and perceptions of computer delivered assessment; a comparison between computer-based and traditional tests in a 'high-stakes' examination. *Journal of Computer Assisted Learning, 27*(5), 440-447. doi:10.1111/j.1365-2729.2011.00409.x

Fox, R. A., McManus, I. C., & Winder, B. C. (2001). The shortened Study Process Questionnaire: An investigation of its structure and longitudinal stability using confirmatory factor analysis. *British Journal of Educational Psychology, 71*(4), 511-530. doi:10.1348/000709901158659

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences, 111*(23), 8410-8415. doi:10.1073/pnas.1319030111

Frein S. T. (2011). Comparing in-class and out-of-class computer-based tests to traditional paper-and-pencil tests in introductory psychology courses. *Teaching of Psychology, 38*(4), 282-287. doi: 10.1177/0098628311421331

Fuentes-Pardo, J. M., García, A. I., Ramírez-Gómez, Á., & Ayuga, F. (2014). Computer-based tools for the assessment of learning processes in higher education: A comparative analysis. In 8th International Technology, Education and Development Conference Proceedings (pp. 976-984), Valencia, Spain

Gibbs, G. (1999). Using assessment strategically to change the way students learn. In S. Brown & A. Glasner (Eds), Assessment matters in higher education (pp. 41–53) Buckingham: S.R.H.E. and Open University Press

Grün, B., & Zeileis, A. (2009). Automatic Generation of Exams in R. *Journal of Statistical Software, 29*(10), 1-14, doi:10.18637/jss.v029.i10

Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioural Statistics, 33*(2), 204-229. doi: 10.3102/1076998607302636.

Handelsman, M. M., Briggs, W. L., Sullivan, N., & Towler, A. (2005). A measure of college student course engagement. *The Journal of Educational Research, 98*(3), 184-192. doi:10.3200/JOER.98.3.184-192

Harks, B., Klieme, E., Hartig, J., & Leiss, D. (2014). Separating Cognitive and Content Domains in Mathematical Competence. *Educational Assessment,19*(4), 243-266. doi:10.1080 /10627197.2014.964114

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81-112. doi:10.3102/003465430298487

Hattie, J., Biggs, J., & Purdie, N. (1996). Effects of learning skills interventions on student learning: A meta-analysis. *Review of Educational Research, 66*(2), 99-36. doi:10.3102/00346543066002099

Hift, R. J. (2014). Should essays and other open ended type of questions retain a place in written summative assessment in clinical medicine? *BMC Medical Education, 14*(1), 249. doi:10.1186/s12909-014-0249-2.

Hochlehnert, A., Brass, K., Moeltner, A., & Juenger, J. (2011). Does medical students' preference of test format (computer-based vs. paper-based) have an influence on performance?. *BMC Medical Education, 11*(1), 89. doi:10.1186/1472-6920-11-89

Hollingshead, L., & Childs, R. A. (2011). Reporting the percentage of students above a cut-score: the effect of group size. *Educational Measurement: Issues and Practice, 30*(1), 36-43. doi:10.1111/j.1745-3992.2010.00198.x

Hornsby, D. J., & Osman, R. (2014). Massification in higher education: Large classes and student learning. *Higher Education, 67*(6), 711-719. doi:10.1007/s10734-014-9733-1

Huff, K. L., & Sireci, S. (2001). Validity issues in computer-based testing. *Educational Measurement: Issues and Practise, 20*(3), 16-25 . doi: 10.1111/j.1745-3992.2001.tb00066.x

Jensen, J. L., Kummer, T. A., & Godoy, P. D. D. M. (2015). Improvements from a flipped classroom may simply be the fruits of active learning. *CBE-Life Sciences Education, 14*(1), ar5. doi:10.1187/10.1187/cbe.14-08-0129

Jeong, H. (2014). A comparative study of scores on computer-based tests and paper-based tests. *Behaviour & Information Technology, 33*(4), 410-422. doi:10.1080/0144929X.2012.710647

Kahu, E. R. (2013). Framing student engagement in higher education. *Studies in Higher Education, 38*(5), 758-773. doi:10.1080/03075079.2011.598505

Kalogeropoulos, N., Tzigounakis, I., Pavlatou, E.A., & Boudouvis, A.G. (2013). Computer-based assessment of student performance in programing courses. *Computer Applications in Engineering Education, 21*(4), 671-683. doi:10.1002/cae.20512

Karpicke, J. D. (2009). Metacognitive control and strategy selection: deciding to practice retrieval during learning. *Journal of Experimental Psychology: General, 138*(4), 469. doi:10.1037/a0017341

Kerdijk, W., Cohen-Schotanus, J., Mulder, B., Muntinghe, F. L., & Tio, R. A. (2015). Cumulative versus end-of-course assessment: effects on self-study time and test performance. *Medical Education, 49*(7), 709-716. doi:10.1111/medu.12756

Kerdijk, W., Tio, R. A., Mulder, B. F., & Cohen-Schotanus, J. (2013). Cumulative assessment: strategic choices to influence students' study effort. *BMC Medical Education, 13*(1), 172. doi:10.1186/1472-6920-13-172

Ketterlin-Geller, L. R., & Yovanoff, P. (2009). Diagnostic assessments in mathematics to support instructional decision making. *Practical Assessment, Research & Evaluation, 14*(16), 1-11. [no doi available]

Kibble, J. (2007). Use of unsupervised online quizzes as formative assessment in a medical physiology course: effects of incentives on student participation and performance. *Advances in Physiology Education, 31*(3), 253-260. doi:10.1152/advan.00027.2007

Kim, Y. H., & Goetz, E. T. (1993). Strategic processing of test questions: The test marking responses of college students. *Learning and Individual Differences, 5*(3), 211-218. doi: 10.1016/1041-6080(93)90003-B

Kolen, M. J., & Brennan, R. L. (2014). Test equating, scaling, and linking: Methods and practices. New York: Springer-Verlag.

Kostal, J. W., Kuncel, N. R., & Sackett, P. R. (2016). Grade inflation marches on: Grade increases from the 1990s to 2000s. *Educational Measurement: Issues and Practice, 35*(1) 11-20. doi:10.1111/emip.12077

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into Practice, 41*(4), 212-218. doi:10.1207/s15430421tip4104_2

Kuh, G. D., Cruce, T. M., Shoup, R., Kinzie, J., & Gonyea, R. M. (2008). Unmasking the effects of student engagement on first-year college grades and persistence. *The Journal of Higher Education, 79*(5), 540-563. doi:10.1353/jhe.0.0019

Lakens, D. (2017). Equivalence Tests: A practical primer for t tests, correlations and meta-analyses. *Social Psychological and Personality Science, 8*(4), 355-362. doi:10.1177/1948550617697177

Lee G., & Weekaron P. (2001). The role of computer-aided assessment in health professional education: A comparison of student performance in computer-based and paper-and-pen multiple-choice tests. *Psychological Bulletin, 23*(2). doi:10.1080/01421590020031066

Lievens, F. (2013). Adjusting medical school admission: Assessing interpersonal skills using situational judgment tests. *Medical Education, 47*(2), 182-189. doi:10.1111/medu.12089

Lim, E. C., Ong, B. K., Wilder-Smith, E. P., & Seet, R. C. (2006). Computer-based versus pen-and-paper testing: Students' perception. *Annals of the Academy of Medicine Singapore, 35*(9), 599-603. [no doi available]

Liu, O. L. (2011). Outcomes assessment in higher education: Challenges and future research in the context of voluntary system of accountability. *Educational Measurement: Issues and Practice, 30*(3), 2-9. doi:10.1111/j.1745-3992.2011.00206.x

Lugtig, P. (2014). Panel attrition separating stayers, fast attriters, gradual attriters, and lurkers. *Sociological Methods & Research, 43*(4). 669-723. doi:10.1177/0049124113520305

Lugtig, P., Glasner, T., & Boevé, A. J. (2015). Reducing underreports of behaviours in retrospective surveys: The effects of three different strategies. *International Journal of Public Opinion Research. 28*(4). 583-595. doi:10.1093/ijpor/edv032

Luyten, H. (1994). Stability of school effects in Dutch secondary education: The impact of variance across subjects and years. *International Journal of Educational Research, 21*(2), 197-216. doi:10.1016/0883-0355(94)90032-9

Marden, N. Y., Ulman, L. G., Wilson, F. S., & Velan, G. M. (2013). Online feedback assessments in physiology: effects on students learning experiences and outcomes. *Advances in Physiology Education, 37*(2), 192-200. doi:10.1152/advan.00092.2012

Mason, G. S., Shuman, T. R., & Cook, K. E. (2013). Comparing the effectiveness of an inverted classroom to a regular classroom in an upper-division engineering course*. IEEE Transactions on Education, 56*(4), 430-435. doi:10.1109/TE.2013.2249066

Maydeu-Olivares A., Kramp U., García-Forero C., Gallardo-Pujol D., & Coffman D. (2009). The effect of varying the number of response alternatives in rating scales: Experimental evidence from intra-individual effects. *Behaviour Research Methods, 41*(2), 295-308. doi: 10.3758/BRM.41.2.295

McDaniel, M. A., Wildman, K. M., & Anderson, J. L. (2012). Using quizzes to enhance summative-assessment performance in a web-based class: An experimental study. *Journal of Applied Research in Memory and Cognition,1*(1), 18-26. doi:10.1016/j.jarmac.2011.10.001

McDonald, A. S. (2002). The impact of individual differences on the equivalence of computer-based and paper-and-pencil educational assessments. *Computers & Education. 39*(3). 299-312. doi:10.1016/S0360-1315(02)00032-5

McLaughlin, J. E., Griffin, L. M., Esserman, D. A., Davidson, C. A., Glatt, D. M., Roth, M. T., ..., & Mumper, R. J. (2013). Pharmacy student engagement, performance, and perception in a flipped satellite classroom. *American Journal of Pharmaceutical Education, 77*(3), 196. doi:10.5688/ajpe779196

McNulty, J. A., Sonntag, B., & Sinacore, J. (2007). Test-taking behaviours on a multiple-choice exam are associated with performance on the exam and with learning style. *Journal of the International Association of Medical Science Educators, 17*(1), 52-57. [no doi available]

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114*(3), 449-458. doi: 10.1037/0033-2909.114.3.449

Niemiec, C. P., & Ryan, R. M. (2009). Autonomy, competence, and relatedness in the classroom Applying self-determination theory to educational practice. *Theory and Research in Education, 7*(2), 133-144. doi:10.1177/1477878509104318

Nikou, S., & Economides, A. A. (2013). Student achievement in paper, computer/web and mobile based assessment. Proceedings of the 6th Balkan Conference on Informatics (BCI), Greece

Nonis, S. A. & Hudson, G. I. (2006). Academic performance of college students: influence of time spent studying and working. *Education for Business, 81*(1): 151-159. doi:10.3200/JOEB.81.3.151-159

Norman, G. (2010). Likert scales, levels of measurement and the “laws” of statistics. *Advances in Health Sciences Education, 15*(5), 625-632. doi:10.1007/s10459-010-9222-y

Noyes J., Garland, K., & Robbins, L. (2004). Paper-based versus computer-based assessment: Is workload another test mode effect?. *British Journal of Educational Technology, 35*(1), 111–113. doi:10.1111/j.1467-8535.2004.00373.x

Peterson, B. K., & Reider, B. P. (2002). Perceptions of computer-based testing: a focus on the CFM examination. *Journal of Accounting Education, 20*(4), 265-284. doi:10.1016/S0748-5751(02)00015-5

Pierce, R., & Fox, J. (2012). Vodcasts and active-learning exercises in a “flipped classroom” model of a renal pharmacotherapy module. *American Journal of Pharmaceutical Education, 76*(10), 196. doi:10.5688/ajpe7610196

Pintrich, P. R., Smith, D. A., García, T., & McKeachie, W. J. (1993). Reliability and predictive validity of the Motivated Strategies for Learning Questionnaire (MSLQ). *Educational and Psychological Measurement, 53*(3), 801-813. doi:10.1177/0013164493053003024

Prince, M. (2004). Does active learning work? A review of the research. *Journal of Engineering Education, 93*(3), 223-231. doi: 10.1002/j.2168-9830.2004.tb00809.x

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Reckase, M. D., & Xu, J. R. (2014). The Evidence for a Subscore Structure in a Test of English Language Competency for English Language Learners. *Educational and Psychological Measurement, 75*(5), 805-825. doi:10.1177/0013164414554416

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment, 95*(2), 129-140. doi:10.1080/00223891.2012.725437

Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: a systematic review and meta-analysis. *Psychological Bulletin, 138*(2), 353-387. doi:10.1037/a0026838

Ricketts, C., & Wilks, S. J. (2002). Improving student performance through computer-based assessment: Insights from recent research. *Assessment & Evaluation in Higher Education. 27*(5), 27(5), 475-479. doi:10.1080/0260293022000009348

Rijksoverheid (2014). Toetsbesluit PO. Staatsblad. https://www.rijksoverheid.nl/onderwerpen/toelating-middelbare-school/documenten/besluiten/2014/01/20/toetsbesluit-po

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software, 17*(5), 1-25. doi:10.18637/jss.v017.i05

Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R., & Carlstrom, A. (2004). Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychological Bulletin, 130*(2), 261-288. doi:10.1037/0033-2909.130.2.261

Robitzsch, R. (2016). sirt: Supplementary Item Response Theory Models. R package version 1.10-0. http://CRAN.R-project.org/package=sirt

Roediger III, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*(3), 181-210. doi:10.1111/j.1745-6916.2006.00012.x

Roediger III, H. L., Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: long-term improvements from quizzing. *Journal of Experimental Psychology: Applied, 17*(4), 382. doi:10.1037/a0026252

Saint, D. A., Horton, D., Yool, A., & Elliott, A. (2015). A progressive assessment strategy improves student learning and perceived course quality in undergraduate physiology. *Advances in Physiology Education, 39*(3), 218-222. doi:10.1152/advan.00004.2015

Schneider, M. C., & Andrade, H. (2013). Lecturers' and Administrators' Use of Evidence of Student Learning to Take Action: Conclusions Drawn from a Special Issue on Formative Assessment. *Applied Measurement in Education, 26*(3), 159-162. doi:10.1080/08957347.2013.793189

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics, 15*(6), 657–680. doi:10.1007/BF01068419

Schuman, H., Walsh, E., Olson, C., & Etheridge, B. (1985). Effort and reward: The assumption that college grades are affected by quantity of study. *Social Forces, 63*(4), 945-966 doi:10.1093/sf/63.4.945

Schuwirth, L. W., & Van der Vleuten, C. P. (2011). Programmatic assessment: from assessment of learning to assessment for learning. *Medical Lecturer, 33*(6), 478-485. doi:10.3109/0142159X.2011.565828

Sinharay, S., Wan, P., Choi, S. W., & Kim, D. I. (2015). Assessing Individual-Level Impact of Interruptions During Online Testing. *Journal of Educational Measurement, 52*(1), 80-105. doi:10.1111/jedm.12064

Sinharay, S., Wan, P., Whitaker, M., Kim, D. I., Zhang, L., & Choi, S. W. (2014). Determining the Overall Impact of Interruptions During Online Testing. *Journal of Educational Measurement. 51*(4), 419-440. doi:10.1111/jedm.12052

Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement, 47*(2), 150-174. doi:10.1111/j.1745-3984.2010.00106.x

Sinharay, S., Puhan, G., & Haberman, S. J. (2011). An NCME instructional module on subscores. *Educational Measurement: Issues and Practice, 30*(3), 29-40. doi:10.1111/j.1745-3992.2011.00208.x

Strayer, J. F. (2012). How learning in an inverted classroom influences cooperation, innovation and task orientation. *Learning Environments Research, 15*(2), 171-193. doi:10.1007/s10984-012-9108-4

Street, S. E., Gilliland, K. O., McNeil, C., & Royal, K. (2015). The flipped classroom improved medical student performance and satisfaction in a pre-clinical physiology course. *Medical Science Educator, 25*(1), 35-43. doi:10.1007/s40670-014-0092-4

Terzis, V., & Economides, A. A. (2011). The acceptance and use of computer based assessment. *Computers & Education, 56*(4), 1032-1044. doi:10.1016/j.compedu.2010.11.017

Tomes, J. L., Wasylkiw, L., & Mockler, B. (2011). Studying for success: diaries of students' study behaviours. *Educational Research and Evaluation, 17*(1), 1-12. doi:10.1080/13803611.2011.563087

Towns, M. H., & Robinson, W. R. (1993). Student use of test-wiseness strategies in solving multiple-choice chemistry examinations. *Journal of Research in Science Teaching, 30*(7), 709-722. doi:10.1002/tea.3660300709

Tune, J. D., Sturek, M., & Basile, D. P. (2013). Flipped classroom model improves graduate student performance in cardiovascular, respiratory, and renal physiology. *Advances in Physiology Education, 37*(4), 316-320. doi:10.1152/advan.00091.2013

Vaessen, B. E., van den Beemt, A., van de Watering, G., van Meeuwen, L. W., Lemmens, L., & den Brok, P. (2016). Students' perception of frequent assessments and its relation to motivation and grades in a statistics course: a pilot study. *Assessment & Evaluation in Higher Education, 42*(6): 872-886. doi:10.1080/02602938.2016.1204532

Van der Drift, K. D. J., & Vos, P. (1987). Anatomie van een leeromgeving, een onderwijseconomische analyse of universitair onderwijs (Anatomy of a learning environment, an economic analysis of university education). Lisse, Netherlands: Swets and Zeitlinger.

Vermunt, J. D., & Verloop, N. (1999). Congruence and friction between learning and teaching. *Learning and Instruction, 9(3)*, 257-280. doi:10.1016/S0959-4752(98)00028-0

Vermunt, J. D., & Vermetten, Y. J. (2004). Patterns in student learning: Relationships between learning strategies, conceptions of learning, and learning orientations. *Educational Psychology Review, 16*(4), 359-384. doi:10.1007/s10648-004-0005-y

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2008). Comparability of Computer-Based and Paper-and-Pencil Testing in K–12 Reading Assessments A Meta-Analysis of Testing Mode Effects. *Educational and Psychological Measurement, 68*(1), 5-24. doi:10.1177/0013164407305592

Wei, X. & Haertel, E. (2011). The effect of ignoring classroom-level variance in estimating the generalizability of school mean scores. *Educational Measurement: Issues and Practice, 30*(1), 13-22. doi:10.1111/j.1745-3992.2010.00196.x

West, S. G., Duan, N., Pequegnat, W., Gaist, P., Des Jarlais, D. C., Holtgrave, D., … & Mullen, P. D. (2008). Alternatives to the Randomized Controlled Trial. *American Journal of Public Health, 98*(8), 1359–1366. http://doi.org/10.2105/AJPH.2007.124446

Whitelock, D. (2009). Editorial: e-assessment: developing new dialogues for the digital age. *British Journal of Educational Technology, 40*(2), 199-202. doi:10.1111/j.1467-8535.2008.00932.x

Wibowo, S., Grandhi, S., Chugh, R., & Sawir, E. (2016). A Pilot Study of an Electronic Exam System at an Australian University. *Journal of Educational Technology Systems, 45*(1), 5-33. doi:10.1177/0047239516646746

Wiliam, D., & Black, P. (1996). Meanings and consequences: a basis for distinguishing formative and summative functions of assessment? *British Educational Research Journal, 22*(5), 537-548. doi:10.1080/0141192960220502

Zimmerman, B. J. (1990). Self-regulated learning and academic achievement: An overview. *Educational Psychologist, 25*(1), 3-17. doi:10.1207/s15326985ep2501_2

ICO Dissertation Series

In the ICO Dissertation Series dissertations are published of graduate students from faculties and institutes on educational research within the ICO Partner Universities: Eindhoven University of Technology, Leiden University, Maastricht University, Open University of the Netherlands, University of Amsterdam, University of Twente, Utrecht University, VU University Amsterdam, and Wageningen University, and formerly University of Groningen (until 2006), Radboud University Nijmegen (until 2004), and Tilburg University (until 2002). The University of Groningen, University of Antwerp, University of Ghent, and the Erasmus University Rotterdam have been 'ICO 'Network partner' in 2010 and 2011. From 2012 onwards, these ICO Network partners are full ICO partners, and from that period their dissertations will be added to this dissertation series.

List update January, 2018 (the list will be updated every year in January)

329. Wolff, C. (16-02-2016). *Revisiting 'withitness': Differences in teachers' representations, perceptions, and interpretations of classroom management.* Heerlen: Open University of the Netherlands.
330. Kok, E.M. (01-04-2016). *Developing visual expertise; from shades of grey to diagnostic reasoning in radiology.* Maastricht: Maastricht University.
331. De Beer, H.T. (11-05-2016). *Exploring Instantaneous Speed in Grade Five: A Design Research.* Eindhoven: Eindhoven University of Technology.
332. Ebbeler, J. (12-05-2016). *Implementing data use in schools: effects on the professional development of educators and the role of school leaders in data teams.* Enschede: University of Twente.
333. Draaijer, S. (10-06-2016). *Supporting Teachers in Higher Education in Designing Test Items.* Amsterdam: Vrije Universiteit Amsterdam.
334. Bos, L.T. (15-06-2016). *Moving Beyond Words. Supporting Text Processing Using a Situation Model approach.* Amsterdam: Vrije Universiteit Amsterdam.
335. Vrugte, J. ter (16-06-2016). *Serious support for serious gaming.* Enschede: University of Twente.
336. Kock, Z.D.Q.P. (23-06-2016). *Toward physics education in agreement with the nature of science: Grade 9 electricity as a case.* Eindhoven: Eindhoven University of Technology.
337. Trinh Ba, T. (28-6-2016) *Development of a course on integrating ICT into inquiry-based science education.* Amsterdam: Vrije Universiteit Amsterdam.
338. Gerken, M. (29-06-2016). *How do employees learn at work? Understanding informal learning from others in different workplaces.* Maastricht: Maastricht University.
339. Louws, M.L. (06-07-2016) Professional learning: *what teachers want to learn.* Leiden: Leiden University.
340. Geel, M.J.M. van, & Keuning T. (08-07-2016). *Implementation and Effects of a Schoolwide Data-Based Decision Making Intervention: a Large-Scale Study.* Enschede: University of Twente.
341. Bouwer, I.R., & Koster, M.P. (02-09-2016) *Bringing writing research into the classroom: The effectiveness of Tekster, a newly developed writing program for elementary students.* Utrecht: Utrecht University.
342. Reijners, P.B.G. (02-09-2016.) *Retrieval as a Cognitive and Metacognitive Study Technique to Learn from Expository Text.* Heerlen: Open University of the Netherlands.
343. Hubers, M.D. (08-09-2016). *Capacity building by data team members to sustain schools' data use.* Enschede: University of Twente.
344. Hsiao, Y.P. (23-09-2016). *Peer Support to Facilitate Knowledge Sharing on Complex Tasks.* Heerlen: Open University of the Netherlands.
345. Scheer, E.A. (23-09-2016). Data-based decision making put to the test. Enschede: University of Twente.
346. Bohle Carbonell, K. (28-9-2016). *May I ask you….? The influence of Individual, Dyadic, and Network Factors on the Emergence of Information in Exchange Teams.* Maastricht: Maastricht University.
347. Claessens, L.C.A. (30-09-2016). *Be on my side, I'll be on your side: Teachers' perceptions of teacher–student relationships.* Utrecht: Utrecht university.

348. Jansen in de Wal, J. (18-11-2016). *Secondary school teachers' motivation for professional learning.* Heerlen: Open University of the Netherlands.

349. Kock, W.D. de. (24-11-2016). *The effectiveness of hints during computer supported word problem solving.* Groningen: University of Groningen.

350. Oonk, C. (07-12-2016). *Learning and Teaching in the Regional Learning Environment: Enabling Students and Teachers to Cross Boundaries in Multi-Stakeholder Practices'.* Wageningen: Wageningen University.

351. Beckers, J. (09-12-2016). *With a little help from my e-portfolio; supporting students' self directed learning in senior vocational education.* Maastricht: Maastricht University.

352. Osagie, E.R. (14-12-2016) *Learning and Corporate Social Responsibility. A study on the role of the learning organization, individual competencies, goal orientation and the learning climate in the CSR adaptation process.* Wageningen: Wageningen University.

353. Baggen, Y. (13-01-2017). *LLLIGHT 'in' Europe - Lifelong Learning, Innovation, Growth and Human capital Tracks in Europe.* Wageningen: Wageningen University.

354. Wouters, A. (09-02-2017). *Effects of medical school selection. On the motivation of the student population and applicant pool.* Amsterdam: VU Medisch Centrum.

355. Baas, D.M. (01-05-2017). Assessment for Learning: more than a tool. Maastricht: Maastricht University.

356. Pennings, J.M. (04-05-2017). *Interpersonal dynamics in teacher-student interactions and relationships.* Utrecht: Utrecht University.

357. Lans, R.M. (18-05-2017). *Teacher evaluation through observation.* Groningen: University of Groningen.

358. Grohnert, T. (18-05-2017). *Judge/Fail/Learn; enabling auditors to make high-quality judgments by designing effective learning environments.* Maastricht: Maastricht University.

359. Brouwer, J. (22-05-2017). *Connecting, interacting and supporting. Social capital, peer network and cognitive perspectives on small group teaching.* Groningen: University of Groningen.

360. Van Lankveld, T.A.M. (20-06-2017). *Strengthening medical teachers' professional identity. Understanding identity development and the role of teacher communities and teaching courses.* Amsterdam: Vrije Universiteit Amsterdam.

361. Janssen, N. (23-06-2017). *Supporting teachers' technology integration in lesson plans.* Enschede: University of Twente.

362. Tuithof, J.I.G.M. (23-06-2017). *The characteristics of Dutch experienced history teachers' PCK in the context of a curriculum innovation.* Utrecht: Utrecht University.

363. Van Waes, S. (23-06-2017). *The ties that teach: Teaching networks in higher education.* Antwerp: University of Antwerp.

364. Evens, M. (30-06-2017). *Pedagogical content knowledge of French as a foreign language: Unraveling its development.* Leuven: KU Leuven.

365. Moses, I. (07-09-2017). *Student-teachers' commitment to teaching.* Leiden: Leiden University.

366. Wansink, B.G.J. (15-09-2017). *Between fact and interpretation. Teachers' beliefs and practices in interpretational history teaching.* Utrecht: Utrecht University.

367. Binkhorst, F. (20-10-2017). *Connecting the dots. Supporting the implementation of Teacher Design Teams.* Enschede: University of Twente.

368. Stoel, G.L. (14-11-2017). *Teaching towards historical expertise. Developing students' ability to reason causally in history.* Amsterdam: University of Amsterdam.

369. Van der Veen, M. (28-11-2017). *Dialogic classroom talk in early childhood education.* Amsterdam: Vrije Universiteit Amsterdam.

370. Frèrejean, J. (08-12-2017). *Instruction for information problem solving.* Heerlen: Open University of the Netherlands.

371. Rezende Da Cunha Junior, F. (19-12-2017). *Online groups in secondary education.* Amsterdam: Vrije Universiteit Amsterdam.

372. Van Dijk, A.M. (22-12-2017). *Learning together in mixed-ability elementary classrooms.* Enschede: University of Twente.

About the author

In 2006 Anja completed High School at the Ukarumpa International School-Secondary Campus in Papua New Guinea. After this she studied *Educational Sciences* at the University of Utrecht from 2007-2010, including a 6 month exchange to the Australian National University in Canberra, Australia. From 2010 to 2013 she continued her studies at Utrecht University completing two research masters in *Educational Science: Learning in Interaction*, and *Methodology and Statistics for the Social and Behavioural Sciences,* while also working as a teaching and research assistant for Joop Hox, Edith de Leeuw, and Peter Lugtig. From 2013 to 2017 she worked on her PhD *Implementing Assessment Innovations in Higher Education,* at the University of Groningen with Rob Meijer, Roel Bosker, and Casper Albers. In 2017 she started working at the Free University of Amsterdam in the Methods and Applied Biostatistics group of the department of Health Sciences.

Publications

Albers, C. J., Boevé, A. J., & Meijer, R. R. (2015). A critique to Akdemir and Oguz (2008): Methodological and statistical issues to consider when conducting educational experiments. *Computers & Education, 87*, 238-242. doi:10.1016/j.compedu.2015.07.001

Boevé, A., Bronkhorst, L., Endedijk, M. D., & Meijer, P. (2015). Tackling methodological challenges to gain new insight into the complexity of student teacher learning in a dual teacher education program. In V. Donche, S. De Mayer, D. Gijbels, & H. van den Bergh (Eds.), Methodological Challenges in Research on Student Learning. (pp. 27-53). (Methodology and Statistics Series). AntwerpApeldoorn: Garant Publishers.

Boevé, A. J., Meijer, R. R., Albers, C. J., Beetsma, Y., & Bosker, R. J. (2015). Introducing Computer-Based Testing in High-Stakes Exams in Higher Education: Results of a Field Experiment. *PLoS ONE, 10*(12), doi:10.1371/journal.pone.0143616

Boevé, A. J., Meijer, R. R., Bosker, R. J., Vugteveen, J., Hoekstra, R., & Albers, C. J. (2017) Implementing the flipped classroom: an exploration of study behaviour and student performance. *Higher Education, 74*(6), 1015-1032. doi:10.1007/s10734-016-0104-y

de Leeuw, E. D., Hox, J. J., & Boevé, A. (2016). Handling Do-Not-Know Answers: Exploring New Approaches in Online and Mixed-Mode Surveys. *Social Science Computer Review, 34*(1)*,* 116-132. doi:10.1177/0894439315573744

Lugtig, P., Glasner, T., & Boevé, A. J. (2015). Reducing Underreports of Behaviors in Retrospective Surveys: The Effects of Three Different Strategies. *International Journal of Public Opinion Research, 28*(4), 583-595. doi:10.1093/ijpor/edv032

Meijer, R., Boevé, A., Tendeiro, J., Bosker, R., & Albers, C. (2017). The Use of Subscores in Exams in Higher Education: When is this Useful? *Frontiers in Psychology, 8*, 305. doi:10.3389/fpsyg.2017.00305

Meijer, R. R., Niessen, A. S. M., & Boevé, A. J. (2015). Rapporteren van subtestscores in de klinische praktijk: Oppassen met de interpretatie [A cautionary note on the use of subtest scores in clinical practice]. *De Psycholoog, 50,* 35-41.

Vugteveen, J., Boevé, A. & Hoekstra, R, (2018). The struggles of a field study: Studying the effectiveness of a flipped classroom. SAGE Research Methods Cases. doi:10.4135/9781526438188

Submitted

Boevé, A. J., Meijer, R. R., Beldhuis, H. J. A., Bosker, R. J., & Albers, C. J. *Natural Variation in Grades and its Implications for Assessing the Effectiveness of Educational Innovations in Higher Education.*

Boevé, A. J., Albers, C. J., Bosker, R. J., Meijer, R. R., & Tendeiro, J. N. *Implementing Practice Tests in Psychology Education.*

Evers, C., Dingemans, A., Junghans, A., & Boevé, A. J. *Feeling Bad Or Feeling Good, Does Emotion Affect Your Consumption Of Food? A Meta-Analysis Of The Experimental Evidence*